Stars and Stellar Processes Lecture Notes

Mike Guidry

This document summarizes the first edition of *Stars and Stellar Processes* by Mike Guidry (Cambridge University Press, 2019) in a format suitable for presentation. Sources and references for the material contained here may be found in that book. Problem references are to problems at the ends of chapters. If the problem number is followed by *** (e.g., Problem 4.21 ***), the problem is solved in the *Student Solutions Manual for Stars and Stellar Processes*. This manual may be found in the Solutions link from the class homepage, and is available from the Cambridge University Press website for the book.

Contents

Part I

Stellar Structure

Chapter 1

Some Properties of Stars

There are no lecture notes for this chapter because I don't cover it in class. Instead I assign it (with some homework problems) for students to read as a concise review of introductory astronomy.

Chapter 2

The Hertzsprung–Russell Diagram

There are no lecture notes for this chapter because I don't cover it in class. Instead I assign it (with some homework problems) for students to read as a concise review of introductory astronomy.

Chapter 3

Stellar Equations of State

Our fundamental initial task in astrophysics is to understand the *structure of stars*. At a minimum, this will require

- A set of equations describing the *behavior of stellar matter in gravitational fields.*
- A set of equations governing
  - *energy production* by thermonuclear reactions and
  - the associated *compositional changes*.
- A set of equations describing *energy transport* from the energy-producing regions deep in the star to the surface.
- *Equations of state* that
  - carry *information about the microscopic physics* of the star and that
  - *relate thermodynamic variables* to each other.

These equations may be coupled in highly non-trivial ways.

Example: Hydrodynamics is influenced by thermonuclear energy production and thermonuclear processes are in turn strongly dependent on variables like temperature and density controlled by the hydrodynamical evolution.

- The full problem will correspond to a set of
  - *coupled*,
  - *non-linear*,

  partial differential equations that can only be solved by large-scale numerical computation.
- In many cases assumptions are justified that allow simpler solutions illustrating many basic stellar features.
- We begin the discussion by considering equations of state.

Equation of State: A *relationship among thermodynamic variables* for a system that

- contains information *beyond what is known from thermodynamics* alone, and is
- often based on *microscopic structure input* from nuclear, atomic, or particle physics.

A *general equation of state* is of the form

$$P = P(T, \rho, X_i, \ldots),$$

where

- P is the *pressure*,
- T is the *temperature*,
- ρ is the *density*,
- the X_i are *concentrations variables* for species i,

and so on.

The preceding equation is intended to be highly schematic at this point: an equation of state

- can take many forms, and
- it need not even be specified analytically.

Example: Equations of state for large-scale astrophysics simulations may be specified by interpolation in tables constructed numerically.

- Fortunately, *relatively simple equations of state suffice* for many (not all!) applications in astrophysics.

3.1 The Pressure Integral

Except possibly at extremely high densities, we are primarily concerned with *equations of states for gases* in astrophysics.

If *quantum effects can be neglected* the pressure in a gas may be expressed in terms of the *pressure integral*:

$$P = \frac{1}{3}\int_0^\infty vpn(p)dp,$$

where

- v is the *velocity*,
- p is the *momentum*,
- $n(p)$ is the *number density* of particles with momentum in the interval p to $p+dp$.

This formula represents a very general result that is *valid for gas particles with any velocity*, up to and including $v = c$.

3.2 Ideal Gases

If the particles in a gas

- *interact weakly* enough,
- the gas obeys the *ideal gas equation of state*,

which may be expressed in a *variety of equivalent forms*:

$$\underbrace{P = nkT = \frac{N}{V}kT = \rho\frac{kT}{\mu}}_{\textit{Forms we will use}} = \frac{NM_u}{V}RT,$$

where

- P is the *pressure*,
- n is the *number density* of gas particles,
- V is the *volume*,
- $N = nV$ is the *number of particles* contained in volume V,
- the *Boltzmann constant* is k and the *temperature* is T,
- the *number of moles* in the gas volume V is NM_u,
- the *universal gas constant* is $R = kN_A$ (where *Avogadro's number* is $N_A = M_u^{-1}$, with M_u the *atomic mass unit*),
- $\mu = \rho/nM_u$ is the *mass for a gas particle* and the *mass density* is ρ.

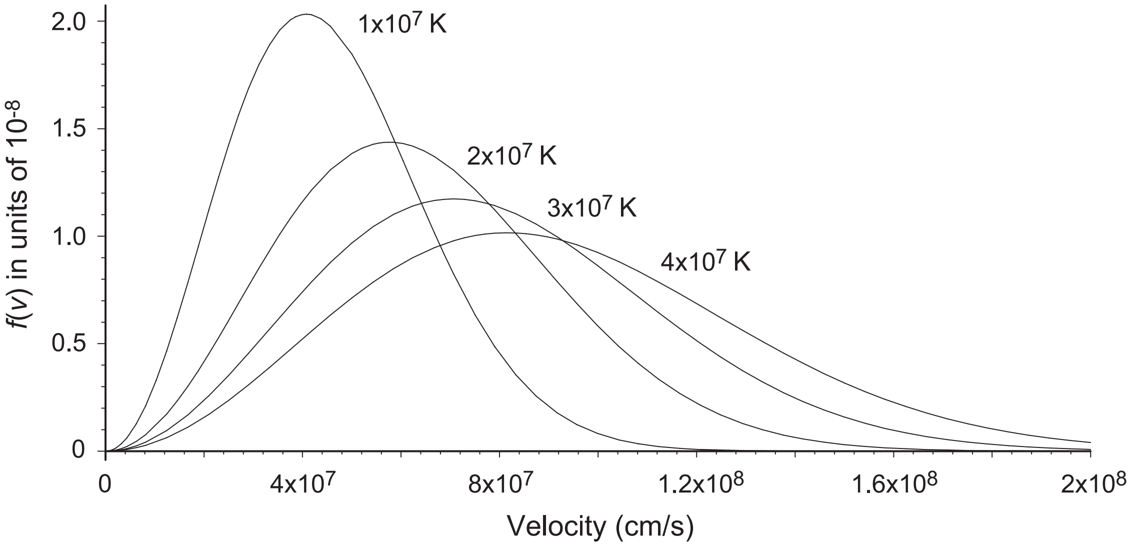


Figure 3.1: Maxwell velocity distribution for hydrogen gas at various temperatures.

The ideal gas equation follows from the *more general pressure integral*

$$P = \frac{1}{3}\int_0^\infty vpn(p)dp,$$

evaluated specifically for a *Maxwellian velocity distribution*,

$$n(p)dp = \frac{4\pi np^2dp}{(2\pi mkT)^{3/2}}e^{-p^2/2mkT}.$$

The Maxwell velocity distribution for hydrogen gas at various temperatures is illustrated in Fig. 3.1.

For an ideal gas the *internal energy* U is given by

$$U = \int_0^T C_V(T)\, dT,$$

where the *heat capacity at constant volume* $C_V(T)$ is

$$C_V(T) \equiv \left(\frac{\partial U}{\partial T}\right)_V = T\left(\frac{\partial S}{\partial T}\right)_V,$$

S is the entropy, and the *first law of thermodynamics,*

$$dU = \delta Q - PdV = T\, dS - PdV,$$

has been used. The *energy density* u is given by

$$u = \frac{U}{V} = \frac{1}{V}\int_0^T C_V\, dT,$$

and the *specific energy* (energy density per unit mass) is

$$\varepsilon = \frac{u}{\rho} = \frac{U}{\rho V} = \frac{1}{\rho V}\int_0^T C_V\, dT.$$

For the special case of a *monatomic, nonrelativistic, ideal gas,*

$$C_V = \frac{3}{2}Nk \qquad U = C_V T = \frac{3}{2}NkT.$$

Expressing the internal energy in differential form,

$$U = \int_0^T C_V(T)\,dT \quad \longrightarrow \quad dU = C_V(T)dT,$$

introducing the *heat capacity at constant pressure* C_P,

$$C_P \equiv \left(\frac{\partial U}{\partial T}\right)_P = T\left(\frac{\partial S}{\partial T}\right)_P,$$

and using the *first law (of thermodynamics)*

$$dU = T\,dS - PdV,$$

we have *for an ideal gas*

$$C_P = C_V + Nk.$$

The *adiabatic index* γ is defined by

$$\gamma \equiv \frac{C_P}{C_V},$$

- *For an ideal gas* the heat capacities are *independent of temperature* and if the gas is *monatomic*,

$$\gamma = \frac{C_P}{C_V} = \frac{C_V + Nk}{C_V} = \frac{\frac{3}{2}Nk + Nk}{\frac{3}{2}Nk} = \frac{5}{3}.$$

Later we will see that γ for an ideal gas is related to the *number of degrees of freedom* per particle.

- The relationship between the pressure P and energy density u for an ideal gas may be expressed in terms of γ:

$$P = (\gamma - 1)u.$$

- This equation may be used to *define an effective adiabatic index* γ for the general case, but *only in the ideal gas limit* is $\gamma = C_P/C_V$.
- The *adiabatic speed of sound in an ideal gas* is given by

$$v_{\mathrm{s}} = \sqrt{\gamma P/\rho},$$

where ρ is density and P is pressure.

3.3 Average Molecular Weights in the Gas

We are concerned with gases consisting of more than one atomic species that may be partially or totally ionized.

- For example, the gas in a star may contain
  - hydrogen atoms and ions,
  - helium atoms and ions,
  - various heavier elements as atoms or ions, and
  - the electrons produced by the ionization.
- In many cases we can treat these mixtures as a single gas with an effective molecular weight.

Example: If density is low enough, a mixture of

- hydrogen ions,
- fully-ionized helium ions, and
- electrons produced by the ionization

will behave as three ideal gases, each contributing a partial pressure to the total pressure *(Dalton's law of partial pressures).*

Then we can treat the system as *a single gas with an effective molecular weight* representing the relative contributions of each individual gas to the system properties.

3.3.1 Concentration Variables

The *mass density* ρ_i of a species i is given by

$$\rho_i = n_i A_i M_u = n_i \frac{A_i}{N_\mathrm{A}},$$

where

- A_i is the *atomic mass number* of species i,
- M_u is the *atomic mass unit*,
- n_i is the *number density* of species i,
- $N_\mathrm{A} = 1/M_u$ is *Avogadro's number*.

Let's introduce the *mass fraction* X_i of species i by

$$X_i \equiv \frac{\rho_i}{\rho} = \frac{n_i A_i M_u}{\rho} = \frac{n_i A_i}{\rho N_\mathrm{A}},$$

where

- ρ is the *total mass density*.
- The label i may refer to *ions, atoms, or molecules*.
- The *mass fractions* sum to unity: $\sum X_i = 1$.

We also will use the *abundance* Y_i,

$$Y_i \equiv \frac{X_i}{A_i} = \frac{n_i}{\rho N_\mathrm{A}}.$$

(NOTE: Generally, the sum of the Y_i will not be unity.)

3.3.2 Total Mean Molecular Weight

As shown in book Appendices, if radiation is ignored the average mass of a gas particle (atoms, ions, electrons) is

$$\mu = \left(\sum_i (1 + y_i Z_i) Y_i \right)^{-1} ,$$

where in this equation

- The sum is over *isotopic species* i.
- y_i is the *fractional ionization* of the species i, with
  - $y_i = 0$ for *no ionization* and
  - $y_i = 1$ for *complete ionization*.
- Z_i is the *atomic number* for isotopic species i.
- Y_i is the *abundance* of isotopic species i.

In very hot stars the *momentum and energy density carried by photons* is non-trivial and we will see later that this further modifies the mean molecular weight of the gas.

We have *replaced the actual gas* (a mixture of electrons and different atomic, possibly molecular, and ionic species) with a gas containing a *single kind of fictitious particle having an effective mass* μ (often termed the *mean molecular weight*).

Example

1. For a *completely ionized gas of atomic hydrogen* there is a single ionic species and $y_i = Z_i = Y_i = 1$. Thus

$$\mu = \frac{1}{\sum_i (1 + y_i Z_i) Y_i} = \frac{1}{(1+1) \times 1} = \frac{1}{2}\,\text{amu}.$$

   This is just the *average mass of a particle in a gas having equal numbers of protons and electrons*, if we neglect the mass of the electrons relative to the protons.

2. The *composition of many white dwarfs* may be approximated by a completely ionized gas consisting of *equal parts* ^{12}C *and* ^{16}O *by mass*. The mass fractions are $X_{12\text{C}} = X_{16\text{O}} = 0.5$, so the abundances are

$$Y_{12\text{C}} = \frac{X_{12\text{C}}}{12} = \frac{0.5}{12} = 0.04167 \qquad Y_{16\text{O}} = \frac{X_{16\text{O}}}{16} = \frac{0.5}{16} = 0.03125.$$

   Therefore,

$$\mu = \frac{1}{(1+1\times 6)\times 0.04167 + (1+1\times 8)\times 0.03125} = 1.745\,\text{amu},$$

   if we assume complete ionization $(y_i = 1)$.

3.3.3 Common Notation

A common shorthand notation:

$$X \equiv X_{\text{hydrogen}} \qquad Y \equiv X_{\text{helium}} \qquad Z \equiv X_{\text{metals}}$$

where *"metals"* refers to the sum of all elements heavier than helium and $X + Y + Z = 1$.

Example: For a typical Pop I star just entering the main sequence (termed a *Zero-Age Main Sequence or ZAMS star*) we find

$$X \simeq 0.7 \qquad Y \simeq 0.3 \qquad Z \simeq 0.02.$$

The metal concentration Z will be less than this in Pop II stars.

3.4 Polytropic Equations of State

An *ideal gas equation of state* (with an effective mean molecular weight μ) is a *realistic approximation for many astrophysical environments*.

- But there are *other possible equations of state* that can play an important role.
- One example is a *polytropic equation of state.*
- A *polytropic process* is defined by the requirement

$$\frac{\delta Q}{\delta T} = c,$$

where

  - δQ is the *change in heat*,
  - δT is the *change in temperature*, and
  - c is a *constant* (don't confuse with the speed of light).

Polytropes have some very useful properties. For example,

- From the first law of thermodynamics and the definition of a polytropic process

  $$\frac{\delta Q}{\delta T} = c,$$

  polytropic processes in ideal gases obey

  $$\frac{dT}{T} = (1-\gamma)\frac{dV}{V},$$

  where the *polytropic* γ is defined by

  $$\gamma \equiv \frac{C_P - c}{C_V - c}.$$

  > The *polytropic* γ reduces to the *ideal gas adiabatic parameter* γ only if the constant $c = 0$.

- You may verify by substitution that the differential equation

  $$\frac{dT}{T} = (1-\gamma)\frac{dV}{V}$$

  has *three classes of solutions*

  $$PV^{\gamma} = C_1 \qquad P^{1-\gamma}T^{\gamma} = C_2 \qquad TV^{\gamma-1} = C_3,$$

  that define *polytropic equations of state,* with C_n being constants.

In astrophysics, the *most common form of a polytropic equation of state* is is

$$P(r) = K\rho^{\gamma}(r) = K\rho^{1+1/n}(r),$$

(Note: $\rho \propto 1/V$, so this is of the form $PV^{\gamma} =$ constant) where the *polytropic index* n is parameterized by

$$n = \frac{1}{\gamma - 1},$$

in terms of the polytropic parameter γ.

1. A polytropic approximation implies physically that the *pressure is independent of temperature,* depending only on density and composition.

2. A polytropic equation of state approximation often simplifies finding solutions for the equations of stellar structure:

   - It *decouples the differential equations* describing hydrostatic equilibrium from those governing energy transfer and the temperature gradients.
   - The decoupled set of equations is generally easier to solve than the original coupled set.

Examples where polytropes are appropriate:

1. For a *completely ionized star, fully mixed by convection* with *negligible radiation pressure*,

   $$P = K\rho^{5/3},$$

   which corresponds to a *polytrope with* $\gamma = \frac{5}{3}$ *and* $n = \frac{3}{2}$. The phenomenological parameter K is *constant for a given star*, but can differ from star to star.

2. For a completely *degenerate gas of nonrelativistic fermions* (defined below)

   $$P = K\rho^{5/3},$$

   again corresponding to a *polytrope with* $\gamma = \frac{5}{3}$ *and* $n = \frac{3}{2}$, but now K *is fixed by fundamental constants*.

3. For a *degenerate gas of ultrarelativistic fermions* (defined below)

   $$P = K\rho^{4/3},$$

   corresponding to a *polytrope with* $\gamma = \frac{4}{3}$ *and* $n = 3$, with K *again fixed by fundamental constants*.

These three polytropic equations of state will be relevant for *homogenous stars* that are completely mixed by convection, *white dwarfs*, and *neutron stars*, respectively.

3.5 Adiabatic Processes

In terms of the heat Q and the entropy S, *adiabatic processes* are defined by the condition that

$$\delta Q = T\,dS = 0.$$

- From the *first law*,

  $$dU = \delta Q - PdV = T\,dS - PdV,$$

  we see that in an adiabatic process the *change in internal energy comes only from* PdV *work:*

  $$dU = -PdV \qquad (\text{since } \delta Q \equiv 0).$$

- Because they do not exchange heat with their environment, *adiabatic processes are fully reversible* ($dS = 0$).

Realistic phenomena in astrophysics are not adiabatic, but many are at least approximately so.

It is standard practice to introduce three *adiabatic exponents* Γ_1, Γ_2, and Γ_3 through

$$\Gamma_1 \equiv \left(\frac{\partial \ln P}{\partial \ln \rho}\right)_S \qquad \frac{\Gamma_2}{\Gamma_2 - 1} \equiv \left(\frac{\partial \ln P}{\partial \ln T}\right)_S$$

$$\Gamma_3 - 1 \equiv \left(\frac{\partial \ln T}{\partial \ln \rho}\right)_S,$$

where the subscripts S remind us that *adiabatic processes occur at constant entropy*, and where the *logarithmic derivatives* are equivalent to

$$\partial \ln A = \frac{\partial A}{A}.$$

This implies equations of state having one of the three forms

$$PV^{\Gamma_1} = c_1 \qquad P^{1-\Gamma_2} T^{\Gamma_2} = c_2 \qquad TV^{\Gamma_3 - 1} = c_3,$$

where the c_n are constants.

Note: From the above definitions,

$$\Gamma_1(\Gamma_2 - 1) = \Gamma_2(\Gamma_3 - 1),$$

so only two of the three Γ_i are independent.

For the *special case of ideal gases*

$$\Gamma_1 \equiv \left(\frac{\partial \ln P}{\partial \ln \rho}\right)_S \qquad \frac{\Gamma_2}{\Gamma_2 - 1} \equiv \left(\frac{\partial \ln P}{\partial \ln T}\right)_S$$

$$\Gamma_3 - 1 \equiv \left(\frac{\partial \ln T}{\partial \ln \rho}\right)_S$$

are equal and equivalent to the ideal gas γ,

$$\Gamma_1 = \Gamma_2 = \Gamma_3 = \gamma \qquad \text{(ideal gas)}.$$

But for more general equations of state Γ_1, Γ_2, and Γ_3 are distinct and carry information emphasizing different aspects of the gas thermodynamics:

1. Because it relates ΔP to $\Delta\rho$, Γ_1 enters into *dynamical properties of the gas* like sound speed.

2. Γ_2 is important for *convective gas motion,* because it relates ΔP to ΔT.

3. Γ_3 *influences the response of the gas to compression,* since it depends on the relationship of ΔT to $\Delta\rho$.

An example of these differences is given below for a mixture of ideal gas and photons in the adiabatic limit.

3.6 Quantum Mechanics and Equations of State

Stellar equations of state reflect *microscopic properties of the gas* in stars.

- A *low-density* gas behaves *classically*,
- A *high-density gas* behaves *quantum mechanically*.

The quantum physics required at high density can be understood in terms of four basic ideas.

1. *deBroglie Wavelength:* The foundation of a quantum description of matter is *particle–wave duality:*
   - Microscopically, particles take on wave properties characterized by a *deBroglie wavelength* $\lambda = h/p$, where p is the momentum and h is Planck's constant.
   - Thus the *particle location becomes fuzzy,* spread out over an interval comparable to $\lambda = h/p$.
2. *Uncertainty Principle:* The *Heisenberg uncertainty principle quantifies the fuzziness* of particle–wave duality:
   - $\Delta p \cdot \Delta x \geq \hbar$, where Δp is the *uncertainty in momentum*, Δx is the *uncertainty in position*, and $\hbar \equiv h/2\pi$.
   - $\Delta E \cdot \Delta t \geq \hbar$, where ΔE is the *energy uncertainty* and Δt is the *time uncertainty* for the energy measurement.

3. *Quantum Statistics:* All elementary particles may be classified as either *fermion* or *bosons*. These classifications indicate how aggregates of elementary particles behave.

   - *Fermions* (such as electrons, or neutrons and protons if we neglect their internal quark and gluon structure) obey *Fermi–Dirac statistics.*
     - The most notable consequence is the *Pauli exclusion principle:* no two fermions can have an identical set of quantum numbers.
     - Elementary particles of *half-integer spin* are fermions.
   - *Bosons* (photons are the most important example for our purposes) obey *Bose–Einstein statistics.*
     - Unlike for fermions, there is *no restriction on how many bosons can occupy the same quantum state*.
     - Elementary particles of *integer spin* are bosons.
   - *Matter is made from fermions* (electrons, protons, neutrons, . . .).
   - *Forces* are mediated by *exchange of bosons*.

   Example: *Electromagnetic forces* result from exchange of photons (which are *bosons*) between charged particles (which can be *fermions* or *bosons*).

4. ***Degeneracy:*** The *exclusion principle* implies that in a many-fermion system *each fermion must be in a different quantum state*.

    - Thus the lowest-energy state results from filling energy levels from the bottom up.
    - *Degenerate matter* corresponds to a many-fermion state in which
        - all the *lowest energy levels are filled* and
        - all the *higher-energy states are empty*.
    - Degenerate matter
        - occurs frequently at *high densities* and
        - has a very *unusual equation of state*.

The equation of state for degenerate matter has a number of consequential implications for astrophysics.

3.7 Equations of State for Degenerate Gases

Degenerate equations of state play an important role in a variety of astrophysical applications. For example,

- In *white dwarf stars* the *electrons are highly degenerate*.
- In *neutron stars* the *neutrons are highly degenerate*.

Let us look at this in a little more detail for the case of *degenerate electrons*.

- We first demonstrate that (as a consequence of quantum mechanics)
  - *most stars are completely ionized over much of their volume* because
  - ionization can be induced by *sufficiently high pressure*, even at low temperature.

This implies the possibility of producing a (relatively) *cold gas of electrons,* which is the necessary condition for a degenerate electron equation of state.

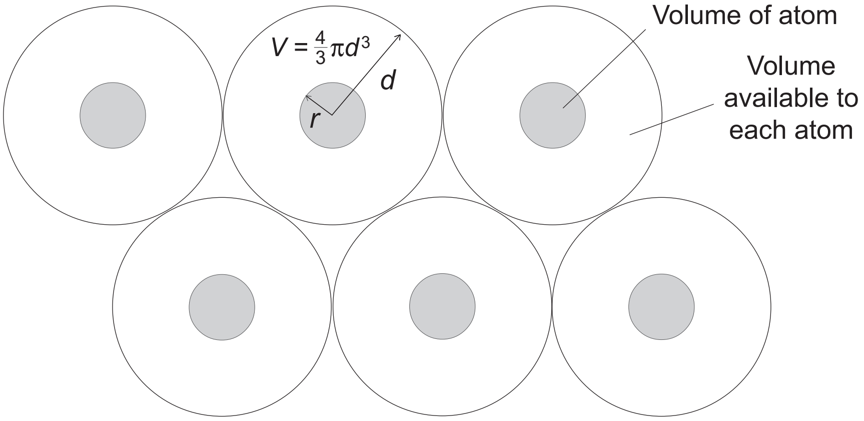
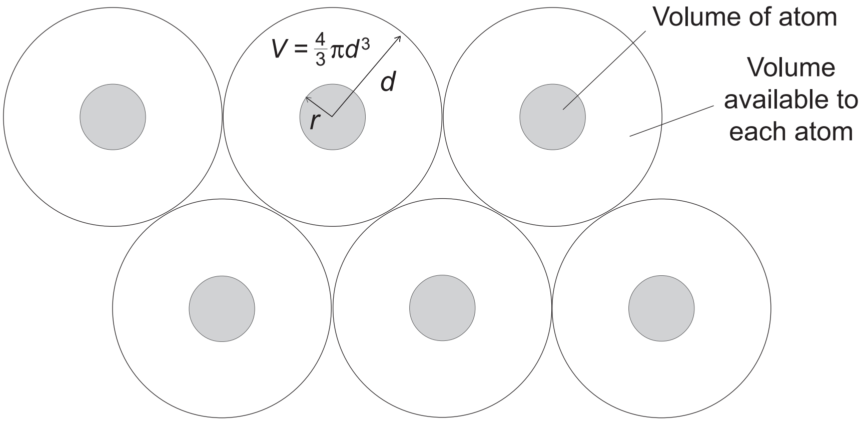


Figure 3.2: Schematic illustration of average atomic spacing in dense stellar matter. These are slices of 3-dimensional spherical volumes

3.7.1 Pressure Ionization

Consider the schematic diagram shown in Fig. 3.2, where

- *Atoms* occupy the *darker spheres* of radius r.
- The *average spacing between atoms* is represented in terms of the *lighter spheres* with radius d.
- To illustrate simply, we assume for that the stellar material consists only of
  - *ions of a single species* and
  - *electrons* produced by ionizing that species.

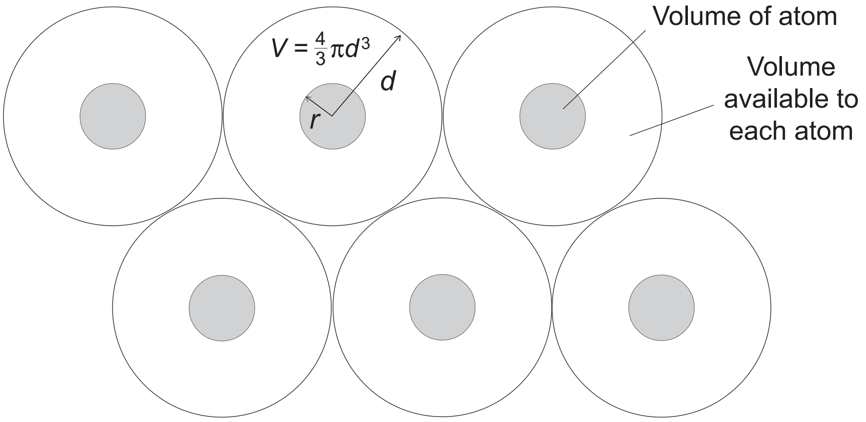


- Electrons in the atoms obey *Heisenberg uncertainty relations* of the form
$$p \cdot \Delta x \geq \hbar,$$
where we've made the usual rough estimate $\Delta p \sim p$.

- Taking an *average volume per electron* of $V_0 \simeq (\Delta x)^3$, the uncertainty relation becomes
$$p \geq \hbar / V_0^{1/3}.$$

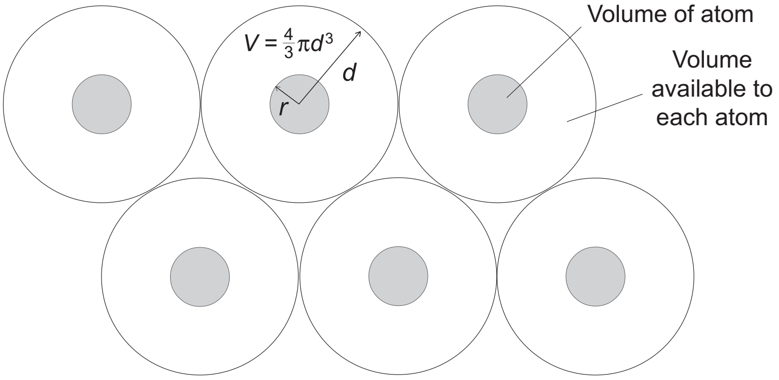
> *The uncertainty principle produces ionization* when the effective volume of the atoms becomes *too small to confine the electrons*.

- The volume per electron V_0 and volume per ion V_i are related by $ZV_0 = V_i$, since there are Z electrons per ion. Thus
$$p \geq \hbar / V_0^{1/3} \quad \longrightarrow \quad p \geq \hbar Z^{1/3} / V_i^{1/3}.$$



- From atomic physics the *radius of an atom* may be approximated by $r \simeq a_0 Z^{-1/3}$, where $a_0 = 5.3 \times 10^{-9}$ cm is the *Bohr radius*.

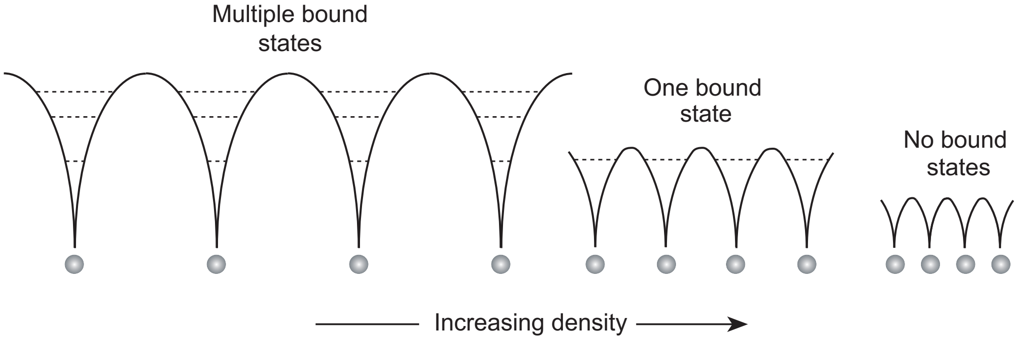
- If the star is composed entirely of an element with atomic number Z and mass number A, there are Z electrons in each sphere of radius d and the *average number density of electrons* is
$$n_e = \frac{Z}{\frac{4}{3}\pi d^3},$$
which may be solved to give the *average spacing of electrons* d,
$$d \simeq \left(\frac{3Z}{4\pi n_e}\right)^{1/3}.$$



- Provided that

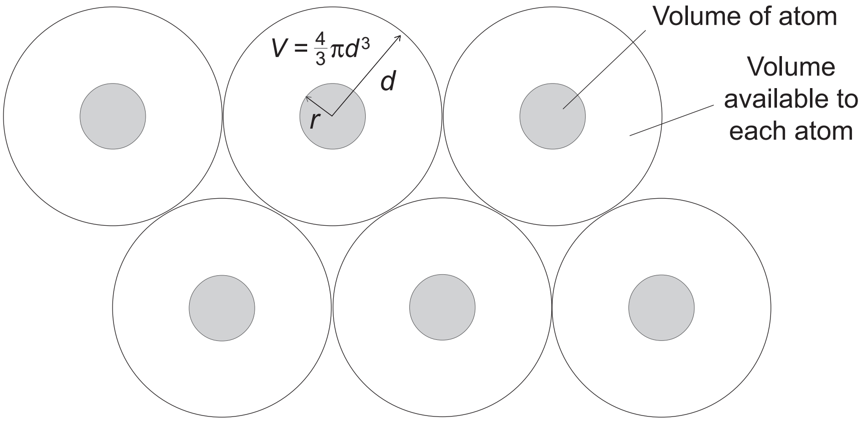
$$d = \left(\frac{3Z}{4\pi n_{\mathrm{e}}}\right)^{1/3} < r,$$

we may expect *pressure ionization,* as illustrated below:



With increasing density fewer locally bound states are possible until none remain and electrons are all ionized.

Thus, *high density can cause complete ionization*, even at zero temperature.



- Since there are A nucleons in each volume of radius d in the above figure, the mass density ρ is

$$\rho = \frac{Am_u}{\frac{4}{3}\pi d^3},$$

and requiring that $d \simeq r \simeq a_0 Z^{-1/3}$ defines a *critical density*

$$\rho_{\text{crit}} \simeq \frac{ZAm_u}{\frac{4}{3}\pi a_0^3}.$$

For densities greater than this there will be *complete pressure ionization,* irrespective of the temperature.

Table 3.1: Critical pressure-ionization densities

Element	(Z,A)	Density ($\mathrm{g\,cm^{-3}}$)
Hydrogen	$(1,1)$	3.2
Helium	$(2,4)$	26
Carbon	$(6,12)$	230
Oxygen	$(8,16)$	410
Iron	$(26,56)$	4660

- The density condition for ionization

$$\rho_{\mathrm{crit}} \simeq \frac{ZAm_u}{\frac{4}{3}\pi a_0^3}$$

  is *satisfied rather easily*.

- Consider *hydrogen gas*: $Z = A = 1$ gives a critical density of $3.2\,\mathrm{g\,cm^{-3}}$, only a factor of ~ 3 larger than for water.

- *Critical pressure ionization densities* for some representative gases are summarized in Table 3.1.

- These may be compared with *actual densities* of

  $\sim 150\,\mathrm{g\,cm^{-3}}$ for the *center of the Sun*,

  $\sim 10^4$–$10^6\,\mathrm{g\,cm^{-3}}$ for a *C–O white dwarf*,

  $\sim 10^9\,\mathrm{g\,cm^{-3}}$ for the iron core of a *massive star*.

The *Saha equations* (describing *thermal ionization*) are not reliable inside stars, where atoms are ionized by *both temperature and pressure*.

By the Saha equations $\sim 24\%$ of the hydrogen in the core of the Sun should be neutral.

- However, comparison of the preceding table with properties of the solar interior indicates that the density is sufficiently high to pressure ionize hydrogen over the inner 40% of the Sun.
- Note that increased pressure favors pressure ionization but disfavors thermal ionization because of increased ion–electron recombination.

Between thermal and pressure effects, *much of the solar interior is entirely ionized,* in contrast to what we would expect from the Saha equations alone.

3.7.2 Classical and Quantum Gases

Let us now consider the distinction between a *classical gas* and a *quantum gas*, and the corresponding implications for stellar structure.

(1) *Identical fermions* are described statistically in quantum mechanics by the *Fermi–Dirac distribution*

$$f(\varepsilon_p) = \frac{1}{\exp[(\varepsilon_p - \mu)/kT] + 1} \qquad \text{(Fermi–Dirac)},$$

where ε_p is given by

$$\varepsilon_p = mc^2 + \frac{p^2}{2m} \qquad \text{(nonrelativistic)},$$
$$\varepsilon_p^2 = p^2c^2 + m^2c^4 \qquad \text{(relativistic)},$$

and the *chemical potential* μ is introduced by adding to the first law of thermodynamics a term accounting for a possible *change in the particle number* N:

$$dU = TdS - PdV + \mu dN,$$

where T, P, and μ are taken as the macroscopic thermodynamical variables for the gas and S is the entropy.

(2) *Identical bosons* are described by the *Bose–Einstein distribution*

$$f(\varepsilon_p) = \frac{1}{\exp[(\varepsilon_p - \mu)/kT] - 1} \qquad \text{(Bose–Einstein)}.$$

We shall consider a gas to be a *quantum gas* if it is described by one of the distributions

$$f(\varepsilon_p) = \frac{1}{\exp[(\varepsilon_p - \mu)/kT] + 1} \qquad \text{(Fermi–Dirac)},$$

$$f(\varepsilon_p) = \frac{1}{\exp[(\varepsilon_p - \mu)/kT] - 1} \qquad \text{(Bose–Einstein)}.$$

and a *classical gas* if the condition

$$e^{(mc^2-\mu)/kT} \gg 1$$

is fulfilled. If the gas is classical,

- The states of lowest energy have $\varepsilon_p \sim mc^2$.
- For fermions or bosons the distribution function becomes well approximated by *Maxwell–Boltzmann statistics*,

  $$f(\varepsilon_p) = e^{-(\varepsilon_p-\mu)/kT} \qquad \text{(Maxwell–Boltzmann)},$$

  where generally $f(\varepsilon_p) \ll 1$ for the classical gas.
- Thus, in a classical gas
  - the lowest energy states are scarcely occupied,
  - the Pauli principle plays little role, and
  - the gas obeys Maxwell–Boltzmann statistics.

We now demonstrate that

- the conditions for forming a classical gas are equivalent to a constraint on the density such that
- the deBroglie wavelength of the particles in the gas is considerably less than the average interparticle spacing in the gas.

Non-Relativistic Classical and Quantum Gases:

Let us introduce a *critical (number) density* n_c

$$n_c \equiv \left(\frac{2\pi mkT}{h^2}\right)^{3/2} = \frac{(2\pi)^{3/2}}{\lambda^3},$$

where the *deBroglie wavelength* λ for non-relativistic particles is given by

$$\lambda = \frac{h}{p} \simeq \left(\frac{h^2}{mkT}\right)^{1/2}$$

(assuming $p = (2mE)^{1/2} \sim (mkT)^{1/2}$). The *number of gas particles* is

$$N = \int_0^\infty f(\varepsilon_p)g(p)\,dp,$$

where the *integration measure* is (See Box 3.7 in book)

$$g(p)dp = g_s \frac{V}{h^3} 4\pi p^2 dp,$$

with p the momentum and $g_s = 2j+1 = 2$ the *spin degeneracy factor* for electrons. Substituting the *Maxwell–Boltzmann distribution* (classical limit of Fermi–Dirac) with a nonrelativistic energy for $f(\varepsilon_p)$,

$$f(\varepsilon_p) = e^{-(\varepsilon_p - \mu)/kT} \qquad \varepsilon_p = mc^2 + \frac{p^2}{2m}$$

integrating, and rearranging yields

$$mc^2 - \mu = kT \ln\left(\frac{g_s n_c}{n}\right),$$

where the *number density* n is given by $n = N/V$.

Therefore, the *classical gas condition*

$$e^{(mc^2-\mu)/kT} \gg 1$$

is, by virtue of

$$mc^2 - \mu = kT \ln\left(\frac{g_\mathrm{s} n_\mathrm{c}}{n}\right),$$

equivalent to

$$e^{\ln\left(\frac{g_\mathrm{s} n_\mathrm{c}}{n}\right)} \gg 1,$$

implying that

$$\frac{g_\mathrm{s} n_\mathrm{c}}{n} \gg 1,$$

which means that *at a given temperature* $n \ll n_\mathrm{c}$ *for a classical gas*, since g_s is of order one.

In a classical gas, the actual number density n is *small on a scale set by the critical density* n_c.

The preceding result has an *alternative interpretation*.

- The *average separation between particles* in the gas is

$$d \sim n^{-1/3} \quad \rightarrow \quad 1/n \sim d^3 .$$

- The *condition $n \ll n_c$ defining a classical gas* implies that $1/n \gg 1/n_c$.

- Because

$$n_c = \frac{(2\pi)^{3/2}}{\lambda^3} \quad \rightarrow \quad n_c \sim \lambda^{-3} \quad \rightarrow \quad 1/n_c \sim \lambda^3$$

the condition $n \ll n_c$ is equivalent to requiring that

$$\frac{1}{n} \gg \frac{1}{n_c} \quad \rightarrow \quad d^3 \gg \lambda^3 \quad \rightarrow \quad d \gg \lambda .$$

For a classical gas, the average separation between particles must be be much larger than the average deBroglie wavelength λ for particles in the gas.

- This makes sense conceptually:
  - The *"quantum fuzziness"* of a particle extends over a distance $\sim \lambda$.
  - If particles are separated on average by distances larger than λ, *quantum effects are minimized*.

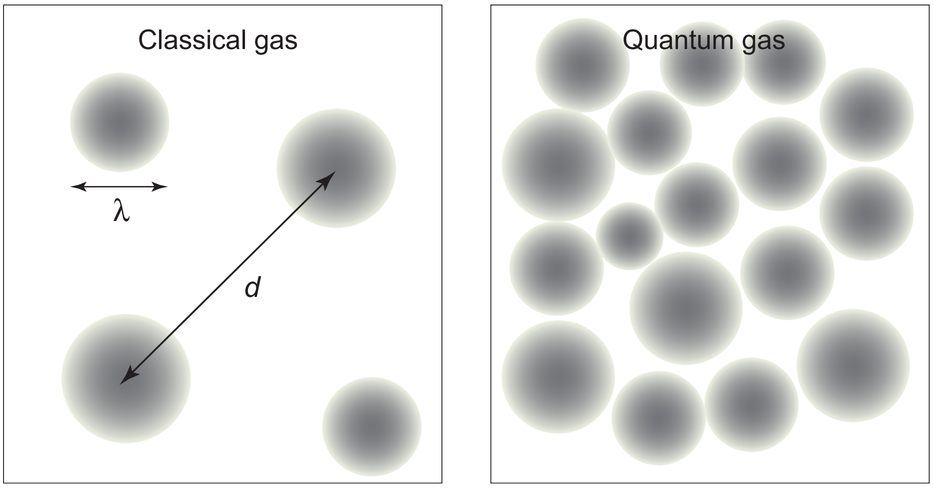


Figure 3.3: Schematic illustration of classical and quantum gases. The width of each fuzzy ball represents the quantum uncertainty in position (not the size) of the particle. In the classical gas (left) the average spacing r between gas particles is much larger than their deBroglie wavelengths λ. In the quantum gas (right) d is comparable to or less than λ. The gas particles have a range of deBroglie wavelengths because they are assumed to have a velocity distribution.

The *schematic relationship between a classical and quantum gas* is illustrated in Fig. 3.3, where

- the size of the spheres represents the *quantum uncertainty in position of a particle, NOT ITS PHYSICAL SIZE,* and
- the spheres have a distribution of sizes because *the deBroglie wavelength depends on velocity* and the gas has a distribution of particle velocities.

Ultrarelativistic Classical and Quantum Gases:

Proceeding in a manner similar to that for the non-relativistic case, for ultrarelativistic particles ($v \sim c$) the rest mass of the particle may be neglected and from

$$\varepsilon_p \simeq kT = \sqrt{m^2c^4 + p^2c^2} \simeq \sqrt{p^2c^2}$$

and

$$f(\varepsilon_p) = e^{-(\varepsilon_p - \mu)/kT},$$

we obtain

$$\mu = -kT \ln\left(\frac{g_s n_c'}{n}\right),$$

where the *relativistic quantum critical density variable* is defined by

$$n_c' = 8\pi\left(\frac{kT}{hc}\right)^3.$$

Hence in the ultrarelativistic case

- The condition that the gas be classical is equivalent to a requirement that $n \ll n_c'$.
- This again is equivalent to requiring that the deBroglie wavelength,

  $$\lambda = \frac{h}{p} \simeq \frac{hc}{kT},$$

  be small compared with the average separation of particles in the gas.

3.7.3 Transition from Classical to Quantum Gas Behavior

We conclude from the preceding results that

> *At high enough density a gas behaves as a quantum rather than classical gas.*

Notice from

$$n_{\mathrm{c}} \equiv \left(\frac{2\pi mkT}{h^2}\right)^{3/2} = \frac{(2\pi)^{3/2}}{\lambda^3}$$

that with increasing gas density:

- The *least massive particles in the gas* will be most prone to a deviation from classical behavior because the critical density is proportional to $n_{\mathrm{c}} \propto m^{3/2}$.
- Thus *photons, neutrinos, and electrons* are most susceptible to such effects.
- The massless *photons never behave as a classical gas* and
- the neutrinos (nearly massless) interact so weakly with matter that they *leave the star unimpeded* when they are produced.
- It follows that in normal stellar environments *the electrons are most susceptible* to a transition from classical to quantum gas behavior.

Presently in the core of the Sun,

- The *average electron number density* is about $6 \times 10^{25}\,\text{cm}^{-3}$.
- The *nonrelativistic critical quantum density* is $n_\text{c} \sim 1.4 \times 10^{26}\,\text{cm}^{-3}$.
- Thus, electrons in the Sun are well approximated by a *dilute classical gas.*

However, the core of the Sun, as for all stars, will contract late in its life as its nuclear fuel is exhausted.

- The approximate relationship between a star's temperature T and radius R is $kT \simeq 1/R$, which implies that

  $$n_\text{c} \equiv \left(\frac{2\pi mkT}{h^2}\right)^{3/2} \simeq \left(\frac{2\pi m}{h^2 R}\right)^{3/2} \simeq R^{-3/2}$$

  and, because the actual number density behaves as $n \sim R^{-3}$, that

  $$\frac{n}{n_\text{c}} \simeq \frac{R^{-3}}{R^{-3/2}} \simeq R^{-3/2}.$$

As the core of the Sun contracts, *eventually the electrons in it will begin to behave as a quantum gas*.

3.7.4 The Degenerate Electron Gas

From the definition

$$n_{\mathrm{c}} \equiv \left(\frac{2\pi mkT}{h^2}\right)^{3/2}$$

the *quantum gas condition* $n \gg n_{\mathrm{c}}$ is equivalent to

$$n \gg \left(\frac{2\pi mkT}{h^2}\right)^{3/2}$$

and thus, solving for kT, we see that $n \gg n_{\mathrm{c}}$ is equivalent to a *temperature constraint,*

$$kT \ll \frac{h^2 n^{2/3}}{2\pi m}.$$

A quantum gas is a cold gas, but *cold on a temperature scale set by the right side of the preceding equation.*

If the density is high enough, a gas could be "cold" while having a temperature of billions of degrees!

- The precise meaning of a cold electron gas is that *the electrons are all concentrated in the lowest available quantum states* consistent with the Pauli principle.
- We say that such a gas is *degenerate*.
- Degenerate gases have much in common with the metallic state in condensed matter.

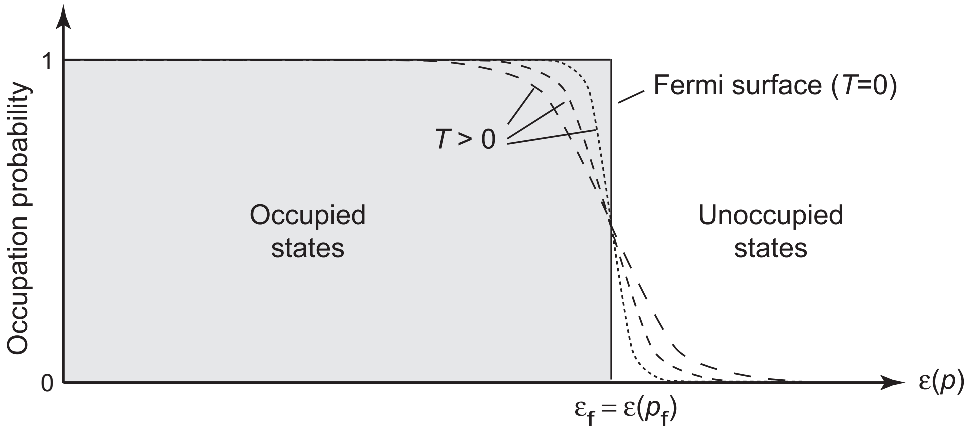


Figure 3.4: The Fermi–Dirac distribution as a function of temperature. Curves with successively shorter dashes represent successively lower temperatures. The solid line defines a step function corresponding to the limit $T \to 0$. This degenerate-gas limit is illustrated further in Fig. 3.5.

As illustrated in Figs. 3.4 and 3.5, in the limit $T \to 0$ *the Fermi–Dirac distribution becomes a step function* in energy space,

$$f_f(\varepsilon_p) = \frac{1}{e^{(\varepsilon_p - \mu)/kT} + 1} \xrightarrow[T\to 0]{} \begin{cases} f(\varepsilon_p) = 1 & \varepsilon_p \le \varepsilon_f \\ f(\varepsilon_p) = 0 & \varepsilon_p > \varepsilon_f \end{cases}$$

- The *chemical potential* μ *at zero temperature* is denoted by ε_f and is termed the *fermi energy*.
- The corresponding value of the momentum is denoted by p_f and is termed the *fermi momentum*.
- Thus, the fermi energy gives the *energy of the highest occupied state* in the degenerate fermi gas.

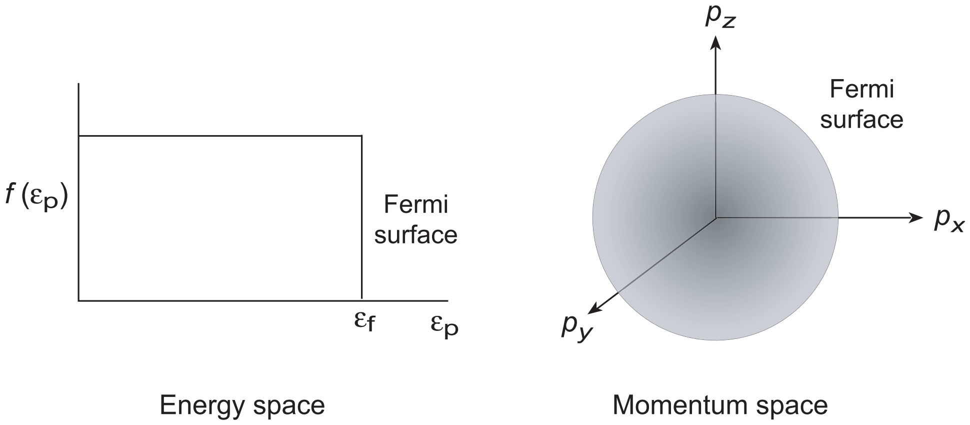


Figure 3.5: The degenerate Fermi gas with its sharp Fermi surface in energy and momentum. In general in condensed matter the Fermi surface may have a more complex shape but it is assumed to be isotropic in momentum for our basic discussion of degenerate gases in stars.

The *density of states as a function of momentum* p is given by $g(p)$ in

$$g(p)dp = g_{\mathrm{s}}\frac{V}{h^3}4\pi p^2 dp,$$

and the *number of electrons in the degenerate gas at zero temperature* is just the number of states with momentum less than the fermi momentum p_{f},

$$\begin{aligned} N &= \int_0^{p_{\mathrm{f}}} g(p)\,dp \\ &= 4\pi V\frac{g_{\mathrm{s}}}{h^3}\int_0^{p_{\mathrm{f}}} p^2\,dp \\ &= \frac{8\pi V}{3h^3}p_{\mathrm{f}}^3, \end{aligned}$$

where $g_{\mathrm{s}} = 2$ has been used for electrons. Solving for the fermi momentum p_{f} and introducing the number density $n = N/V$, we find that *the fermi momentum is determined completely by the electron number density*

$$p_{\mathrm{f}} = \left(\frac{3h^3}{8\pi}\cdot\frac{N}{V}\right)^{1/3} = \left(\frac{3n}{8\pi}\right)^{1/3} h.$$

For a number density n the interparticle spacing is $\sim n^{-1/3}$, implying that the deBroglie wavelength of an electron at the fermi surface, $\lambda = h/p_{\mathrm{f}} \sim n^{-1/3}$, is comparable to the average spacing between electrons.

We may construct the *equation of state* for the degenerate electron gas by evaluating the *internal energy of the gas*.

Let us first do this in the nonrelativistic and then in the ultrarelativistic limits for degenerate electrons.

Nonrelativistic Degenerate Electrons:

In the nonrelativistic limit $p_\mathrm{f} \ll mc$, which implies that

$$n \ll \left(\frac{1}{\lambda_\mathrm{c}}\right)^3 = \left(\frac{mc}{\hbar}\right)^3,$$

where $\lambda_\mathrm{c} \equiv \hbar/mc$ is the *Compton wavelength* for an electron. In this limit the *internal energy* for a degenerate electron gas is

$$U = \int_0^\infty \varepsilon_p f(\varepsilon_p) g(p)\, dp \simeq \underbrace{Nmc^2}_{\text{potential}} + \underbrace{\frac{3N}{10m} p_\mathrm{f}^2}_{\text{kinetic}},$$

where

$$\varepsilon_p = mc^2 + \frac{p^2}{2m} \qquad g(p) = g_\mathrm{s} \frac{V}{h^3} 4\pi p^2 \qquad N = \frac{8\pi V}{3h^3} p_\mathrm{f}^3$$

and $g_\mathrm{s} = 2$ have been used. For a nonrelativistic gas the *pressure is given by $\frac{2}{3}$ of the kinetic energy density;* identifying the second term of

$$U = Nmc^2 + \frac{3N}{10m} p_\mathrm{f}^2,$$

divided by the volume V as the kinetic energy density,

$$P = \frac{2}{3} \times (\text{kinetic energy density}) = \frac{2}{3}\left(\frac{N}{V}\frac{3p_\mathrm{f}^2}{10m}\right)$$

$$= n\frac{p_\mathrm{f}^2}{5m} = \frac{h^2}{5m}\left(\frac{3}{8\pi}\right)^{2/3} n^{5/3} \qquad (\gamma = \tfrac{5}{3} \text{ polytrope}),$$

where $n = N/V$ and $p_\mathrm{f} = (3n/8\pi)^{1/3} h$ have been used.

Example: For a low-mass white dwarf with $\rho \lesssim 10^6\,\mathrm{g\,cm^{-3}}$

- the *electrons are nonrelativistic* and
- the *electron pressure* is given by the $\gamma = \frac{5}{3}$ polytrope implied by the preceding equation,

$$P_\mathrm{e} = \frac{h^2}{5m}\left(\frac{3}{8\pi}\right)^{2/3}\left(\frac{\rho}{m_\mathrm{p}\mu_\mathrm{e}}\right)^{5/3},$$

  with the mean molecular weight μ_e defined through

$$n_\mathrm{e} = \frac{\rho}{m_\mathrm{p}\mu_\mathrm{e}}.$$

As noted earlier, the constant K in the polytropic form

$$P = K\rho^\gamma(r)$$

is *fixed by fundamental constants*.

Ultrarelativistic Degenerate Electrons:

For ultrarelativistic electrons, $n \gg n'_{\rm c}$ implies that

$$n \gg (mc/h)^3.$$

Utilizing the ultrarelativistic limit $\varepsilon_p = pc$ and

$$g(p) = g_{\rm s} \frac{V}{h^3} 4\pi p^2,$$

the internal energy is

$$\begin{aligned} U &= \int_0^\infty \varepsilon_p f(\varepsilon_p) g(p)\, dp \\ &\simeq \frac{8\pi V c}{h^3} \int_0^{p_{\rm f}} p^3\, dp = \tfrac{3}{4} N c p_{\rm f}. \end{aligned}$$

For an ultrarelativistic gas the *pressure is $\frac{1}{3}$ of the kinetic energy density.*

Identifying the kinetic energy density as $U = \frac{3}{4} N c p_{\rm f}$ divided by the volume V, for ultrarelativistic particles

$$\begin{aligned} P &= \frac{1}{3} \times (\text{kinetic energy density}) \\ &= \frac{1}{3} \times \left(\frac{3}{4} c n p_{\rm f} \right) = \frac{hc}{4} \left(\frac{3}{8\pi} \right)^{1/3} n^{4/3} \qquad (\gamma = \tfrac{4}{3} \text{ polytrope}), \end{aligned}$$

where $n = N/V$ and we have used $p_{\rm f} = (3n/8\pi)^{1/3} h$.

Example: For *higher-mass white dwarfs* having $\rho \gtrsim 10^6\,\mathrm{g\,cm^{-3}}$

- the *electrons are highly relativistic* and
- the corresponding *degenerate equation of state* takes the form implied by the preceding equation,

$$P_\mathrm{e} = \frac{hc}{4}\left(\frac{3}{8\pi}\right)^{1/3}\left(\frac{\rho}{m_\mathrm{p}\mu_\mathrm{e}}\right)^{4/3}$$

  which is a *polytrope with* $\gamma = \frac{4}{3}$.

As in the nonrelativistic case, we see that the constant K multiplying $\rho^{4/3}$ is *fixed by fundamental constants*.

3.7.5 Summary: High Gas Density and Stellar Structure

We may identify several important *consequences of high densities* in stellar environments:

- An increase in the gas density above a critical amount *enhances the probability for pressure ionization.*
- This creates a *fully-ionized gas of electrons and ions* irrespective of possible thermal ionization.
- *An increase in the gas density*, by uncertainty principle arguments, *increases the average momentum* of gas particles.
- Thus particles become *more relativistic at high densities*.
- *Increased density raises the fermi momentum.* This, for example, *influences weak interaction processes* in the star.
- *Increase in the gas density decreases the interparticle spacing* relative to the average deBroglie wavelength.
- This makes it more likely that the least massive particles in the system *transition from classical to degenerate quantum gas behavior*.

- Increased density *enhances the strength of the gravitational field* and makes it more difficult to maintain stability of the star against gravitational collapse.
- Higher density also makes it more likely that *general relativistic corrections to Newtonian gravitation become important*.
- Higher density (often implying higher temperature) tends to *change the rates* of thermonuclear reactions and to *alter the opacity* of the stellar material to radiation.
  - The former changes the rate of energy production;
  - the latter changes the efficiency of how that energy is transported in the star.

  Both can have large consequences for stellar structure and evolution.

These *consequences of increased density*

- have *large implications for stellar structure and evolution* because
- all stars are expected to dramatically *increase their central densities during late evolutionary stages*.

From the preceding, a gas can have a pressure that is of *purely quantum-mechanical origin,* independent of its temperature.

- Assume *pressure dominated by non-relativistic electrons* and drop factors like $\frac{1}{2}$.
- For an ideal gas the *average energy of an electron* is $E \sim kT = \frac{1}{2} m_{\mathrm{e}} v^2$, giving an *electron velocity*

  $$v_{\text{thermal}} \simeq \left(\frac{kT}{m_{\mathrm{e}}} \right)^{1/2},$$

  of *purely thermal origin*.
- But *even at zero temperature electrons have a velocity* v_{QM} implied by the *uncertainty principle*, since

  $$p \sim \Delta p \sim \hbar / \Delta x \sim \hbar n_{\mathrm{e}}^{1/3} \quad \rightarrow \quad v_{\mathrm{QM}} \simeq \frac{p}{m_{\mathrm{e}}} \simeq \frac{\hbar n_{\mathrm{e}}^{1/3}}{m_{\mathrm{e}}}.$$

- Thus, there are *two contributions* to the *average velocity of particles* in the gas,

  1. one from the *finite temperature* and
  2. one from purely *quantum effects*,

  with the *thermal contribution vanishing as* $T \rightarrow 0$.

This has much in common with the distinction between

- a *thermal phase transition* (driven by temperature fluctuations that vanish as $T \to 0$), and
- a *quantum phase transition* (driven by quantum fluctations that remain as $T \to 0$).

Such concepts are important in fields like condensed matter physics.

- The *pressure contributed by the thermal motion* is

  $$P_{\text{thermal}} = n_{\text{e}}kT = n_{\text{e}}m_{\text{e}}v_{\text{thermal}}^2$$

  and the *pressure contributed by quantum mechanics* is (up to some constant factors)

  $$P_{\text{QM}} \simeq \frac{\hbar^2}{m_{\text{e}}}n_{\text{e}}^{5/3} = n_{\text{e}}m_{\text{e}}\left(\frac{\hbar n_{\text{e}}^{1/3}}{m_{\text{e}}}\right)^2 = n_{\text{e}}m_{\text{e}}v_{\text{QM}}^2$$

- A *degenerate gas* is one for which $P_{\text{QM}} \gg P_{\text{thermal}}$.
- *Thermal pressure* is proportional to T and density, but *quantum pressure* is independent of T and proportional to a power of density:

  $$P_{\text{thermal}} = n_{\text{e}}kT \qquad P_{\text{QM}} \simeq \frac{\hbar^2}{m_{\text{e}}}n_{\text{e}}^{5/3}.$$

- Therefore, *degeneracy is favored in low-temperature, dense gases*, and
- a gas can have a high *pressure of purely uncertainty-principle origin*, even at $T = 0$.
- Furthermore, *changing T in a degenerate gas will have little effect on the pressure* (as long as changing T does not significantly change the degeneracy).

All of these properties have *profound consequences for stars* when high densities are encountered.

3.8 Equation of State for Radiation

We may view electromagnetic radiation in stars as a gas of *ultrarelativistic massless bosons.*

- The *equation of state for radiation* follows from the energy density and pressure associated with the *Planck frequency distribution*

$$n(\nu)d\nu = \frac{8\pi\nu^2 d\nu}{c^3(e^{h\nu/kT}-1)}.$$

- This yields for the *radiation pressure*,

$$P_{\text{rad}} = \frac{1}{3}aT^4,$$

  where a is the radiation density constant.

- The corresponding *energy density of the radiation field* is

$$u_{\text{rad}} = aT^4 = 3P_{\text{rad}},$$

  implying that $P_{\text{rad}} = \frac{1}{3}u_{\text{rad}}$.

Gravitational Stability and Adiabatic Exponents for Radiation:

As you are asked to show in an exercise,

- adiabatic exponents Γ_1, Γ_2, and Γ_3 for a *pure radiation field* are all equal to $\frac{4}{3}$.
- As we shall see in more detail later, an adiabatic exponent less than $\frac{4}{3}$ generally implies an *instability against gravitational collapse.*

Therefore, *admixtures of radiation contributions* (more generally, of any relativistic component) to pressure often signal *decreased gravitational stability* for a gas.

3.9 Equation of State for Matter and Radiation

For a simple stellar model, it is often a good starting point to assume

- an *ideal gas equation of state for the matter* (provided that the density is not too high) and
- a *blackbody equation of state for the radiation*.

In that case we may write for the pressure P and internal energy U

$$P = \underbrace{\frac{N}{V}kT}_{\text{Ideal gas}} + \underbrace{\frac{aT^4}{3}}_{\text{Radiation}}$$

$$U = uV = \underbrace{C_V T}_{\text{Ideal gas}} + \underbrace{aT^4 V}_{\text{Radiation}} ,$$

where the first term in each case is the contribution of the ideal gas and the second term is that of the radiation.

3.9.1 Mixtures of Ideal Gases and Radiation

For *mixtures of gas and radiation* (common in high-temperature stellar environments),

- it is convenient to define a parameter β that measures the relative contributions of *gas pressure* P_{g} and *radiation pressure* P_{rad} to the *total pressure* P:

$$\beta = \frac{P_{\mathrm{g}}}{P} \qquad 1-\beta = \frac{P_{\mathrm{rad}}}{P} \qquad P = P_{\mathrm{g}} + P_{\mathrm{rad}}.$$

- Therefore
  - $\beta = 1$ implies that *all pressure is generated by the gas.*
  - $\beta = 0$ implies that *all pressure is generated by radiation.*

For all values in between $\beta = 0$ and $\beta = 1$ the pressure receives *contributions from both gas and radiation*.

Example: In a mixture of ideal gas and radiation the *pressure generated by the gas alone* is

$$P_{\mathrm{g}} = nkT = \beta P.$$

Solving this equation for the *total pressure*,

$$P = \frac{nkT}{\beta} = \frac{\rho kT}{\beta\mu} = \frac{NkT}{\beta V},$$

which is of *ideal gas form*. Thus, *mixing radiation with gas*

- gives an *ideal gas equation of state* but
- particles have an *effective mean molecular weight* $\beta\mu$, where μ is the mean molecular weight for the gas alone.

Thus mixtures of ideal gases and radiation may be *treated as modified ideal gases,* but

- normally the relative contribution of radiation and gas to the pressure *varies through the volume of a star*, so
- β is a *local function of position*.

3.9.2 Adiabatic Systems of Gas and Radiation

The preceding discussion of gas and radiation mixtures *depends only on the ideal gas assumption* and the *Planck radiation distribution assumption*.

- Now let's further restrict to *adiabatic processes.*
- Then from
  - the *adiabatic condition* $\delta Q = 0$,
  - the *first law* of thermodynamics, and
  - the *definition* of β,

  at constant entropy (see *Problem 3.16* *** in book)

$$\frac{d\ln T}{d\ln V} = \frac{-(\gamma-1)(4-3\beta)}{\beta+12(\gamma-1)(1-\beta)}.$$

These logarithmic derivatives may then be used to *evaluate the adiabatic exponents*, with the results

$$\Gamma_1 = \frac{d\ln P}{d\ln\rho} = \beta + \frac{(4-3\beta)^2(\gamma-1)}{\beta+12(1-\beta)(\gamma-1)}$$

$$\Gamma_2 = \left(1-\frac{d\ln T}{d\ln P}\right)^{-1} = 1 + \frac{(4-3\beta)(\gamma-1)}{\beta^2+3(\gamma-1)(1-\beta)(4+\beta)}$$

$$\Gamma_3 = 1 + \frac{d\ln T}{d\ln\rho} = 1 + \frac{(4-3\beta)(\gamma-1)}{\beta+12(1-\beta)(\gamma-1)}$$

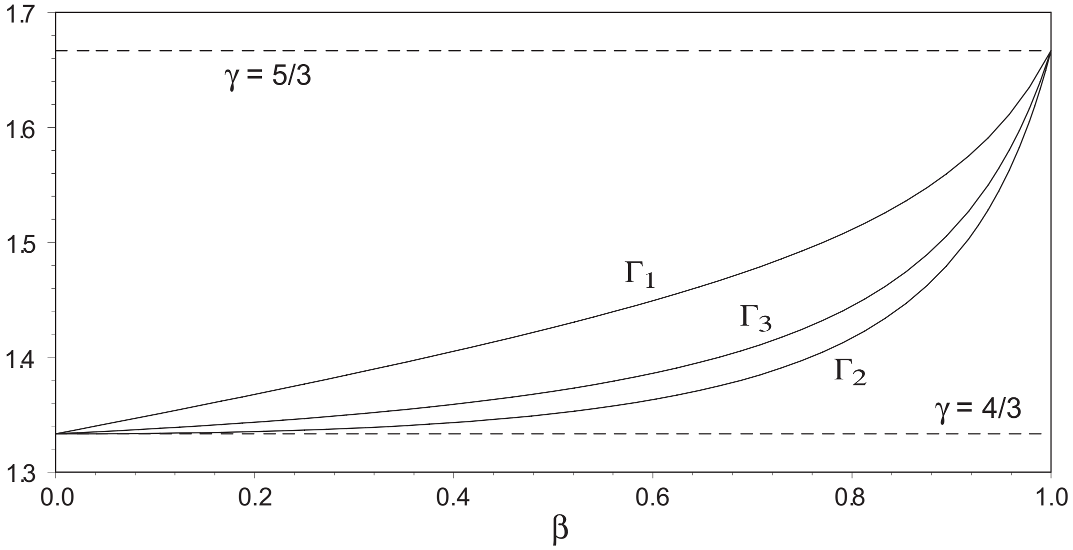


Figure 3.6: Adiabatic exponents in a mixture of ideal gas and radiation.

The adiabatic exponents Γ_1, Γ_2, and Γ_3 are plotted in Fig. 3.6 as a function of the parameter β. Notice the *expected limiting behavior:*

- Assuming $\gamma = \frac{5}{3}$ for a *monatomic ideal gas* and $\beta = 1$ (*no radiation contribution to pressure*) gives

$$\Gamma_1 = \Gamma_2 = \Gamma_3 = \frac{5}{3} \qquad \text{(Monatomic ideal gas)}.$$

- For $\beta = 0$ (*all pressure generated by radiation*) we find

$$\Gamma_1 = \Gamma_2 = \Gamma_3 = \frac{4}{3} \qquad \text{(Pure radiation)}.$$

- For other values of β the *adiabatic exponents are not equal* and lie between $\frac{4}{3}$ and $\frac{5}{3}$.

Chapter 4

Hydrostatic and Thermal Equilibrium

A fundamental property of main sequence stars like our Sun is their *stability over long periods of time*.

- The fossil record indicates that *the Sun has been emitting energy at its present rate for several billion years*, with relatively small variation.
- The key to this stability is that main sequence stars are in a state of *near perfect hydrostatic equilibrium*, where
- the *forces deriving from pressure gradients* produced by thermonuclear fusion and internal heat *almost exactly balance the gravitational forces*.

Thus the starting point for understanding stellar structure is an understanding of *hydrostatic equilibrium* and *departures from that equilibrium*.

4.1 Newtonian Gravitation

The *Newtonian gravitational field* is derived from a *gravitational potential* Φ that obeys the *Poisson equation*, which for spherical symmetry is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right) = 4\pi G\rho .$$

The *gravitational acceleration* is given by

$$g = \frac{\partial\Phi}{\partial r} = \frac{Gm}{r^2},$$

where $m = m(r)$ is the *mass contained within the radius* r. Hence, for spherical geometry

$$\Phi(r) = \int_0^r g\,dr + \text{constant} = \int_0^r \frac{Gm(r)}{r^2}dr + \text{constant}.$$

The constant is fixed by requiring that $\Phi \to 0$ as $r \to \infty$.

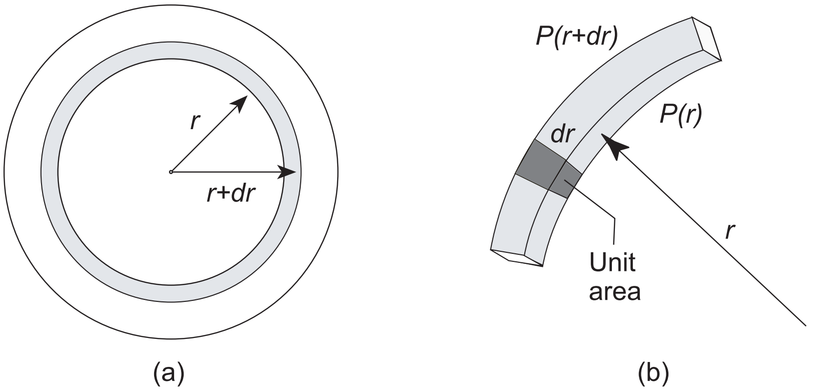


Figure 4.1: Spherical mass shells. In (b) the small shaded volume has height dr and unit area on its inner surface. Therefore its volume is $1 \times dr = dr$ and its mass is $\Delta m = \rho \times 1 \times dr = \rho dr$.

4.2 Conditions for Hydrostatic Equilibrium

The *local gravitational acceleration* at a radius r is given by

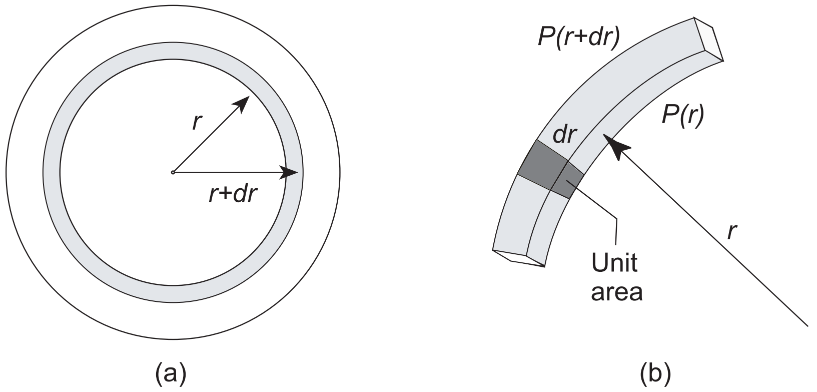
$$g = \frac{\partial \Phi}{\partial r} = \frac{Gm}{r^2},$$

where $m(r)$ is the *mass contained within a radius* r. The mass contained in a *thin spherical shell* is (see Fig. 4.1)

$$dm = m(r + dr) - m(r) = 4\pi r^2 \rho(r) dr.$$

Integrating this from the origin to a radius r yields the *mass function* $m(r)$,

$$m(r) = \int_0^r 4\pi r^2 \rho \, dr.$$



Consider the *total gravitational force* acting on a volume of *unit area* in a *concentric spherical shell* of radius r and depth dr.

- The *magnitude of this force* (per unit area) will be

$$F_{\rm g} = -\rho g(r) dr = -\rho \frac{Gm(r)}{r^2} dr,$$

Negative sign → *directed toward the center of the sphere.*

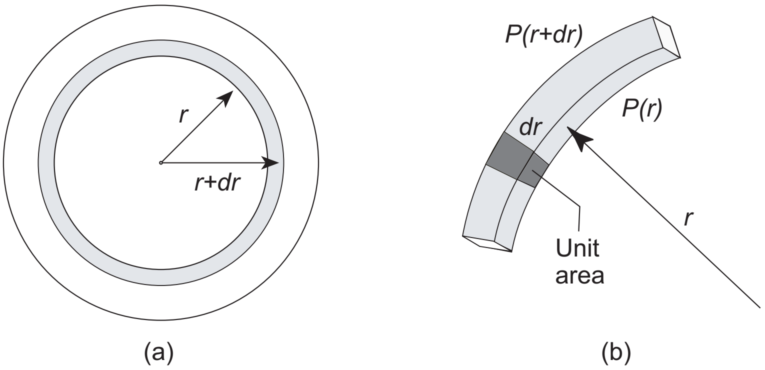
- The *force per unit area* resulting from the pressure difference between r and $r + dr$ is

$$P(r) - P(r+dr) = -\frac{\partial P}{\partial r} dr$$

Negative sign → *directed outward.* ($\partial P / \partial r$ is negative.)

- The *inwardly directed gravitational force* is counterbalanced by a net *outward force arising from the pressure gradient* of the gas and radiation that has a magnitude

$$F_{\rm p} = P(r) - P(r+dr) = -\frac{\partial P}{\partial r} dr.$$



- The *total force* acting on this volume of *unit surface area* is then

$$F = F_{\mathrm{g}} + F_{\mathrm{p}} = -\frac{\partial P}{\partial r}dr - \frac{Gm(r)}{r^2}\rho dr,$$

- by *Newton's 2nd law* the equation of motion is

$$F = ma = \underbrace{\rho dr}_{\text{mass}} \; \underbrace{\frac{\partial^2 r}{\partial t^2}}_{\text{acceleration}} \; ,$$

- This leads to

$$\rho \frac{\partial^2 r}{\partial t^2} = -\frac{\partial P}{\partial r} - \frac{Gm(r)}{r^2}\rho .$$

- For *hydrostatic equilibrium,* the left side vanishes because the acceleration $\partial^2 r/\partial t^2 = 0$ and we obtain

$$\frac{dP}{dr} = -\underbrace{\frac{Gm(r)}{r^2}}_{g}\rho = -g\rho,$$

  where partial derivatives have been replaced with derivatives since we assume no time dependence.

Hydrostatic Equilibrium and Stellar Interiors:

In the equation

$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho = -g\rho,$$

both ρ and $Gm(r)/r^2$ are positive.

1. Thus $dP/dr \leq 0$ and *pressure must decrease outward everywhere* for a gravitating system to be in hydrostatic equilibrium.

   dP/dr *is* always negative *under conditions of hydrostatic equilbrium.*

2. This will in turn imply that density and temperature must increase toward the center of a star.

   The conditions of hydrostatic equilibrium are sufficient to ensure that stars must be

   - much more dense and
   - much hotter

   near their centers than near their surfaces.

The equations

$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho = -g\rho \qquad \text{(Hydro equilibrium)},$$

$$dm = 4\pi r^2 \rho(r) dr \qquad \text{(Mass equation)}.$$

are our first two *equations of stellar structure*.

- They constitute *two equations in three unknowns* (P, m, and ρ as functions of r).
- This system of equations may be *closed by specifying an equation of state* relating these quantities.

Before considering that, we explore some consequences that follow from these equations alone.

4.3 Lagrangian and Eulerian Descriptions

In studying fluid motion, there are two basic computational points of view that we can take.

1. We can fix a grid and watch the fluid flow through the grid; this is called *Eulerian hydrodynamics*.
2. Alternatively, we can construct coordinates that are attached to the mass elements and move with them; this is called *Lagrangian hydrodynamics.*

   > Consider determining the *temperature of the atmosphere* over time either from a *balloon drifting with the wind* or from a *fixed point on the ground*.
   >
   > - The first is a *Lagrangian* point of view, if we imagine the balloon to be tied approximately to the motion of a fixed packet of air.
   > - The second is *Eulerian*, since we observe the air from a fixed point as it flows by.

3. If fluid accelerations can be neglected, Lagrangian and Eulerian descriptions of hydrodynamics reduce to *Lagrangian and Eulerian hydrostatics*.

4.3.1 Lagrangian Formulation of Hydrostatics

Let's illustrate the Lagrangian approach by writing the previous equations with $m(r)$ instead of r as the independent variable.

- The general result for a *change of variables between Eulerian and Lagrangian representations*, $(r,t) \to (m,t)$, is

$$\frac{\partial}{\partial m} = \frac{\partial}{\partial r} \cdot \frac{\partial r}{\partial m} \qquad \left(\frac{\partial}{\partial t}\right)_m = \frac{\partial}{\partial r} \cdot \left(\frac{\partial r}{\partial t}\right)_m + \left(\frac{\partial}{\partial t}\right)_r .$$

  where the subscripts denote variables held constant.

- Apply the first of these to the mass parameter m and use $dm = 4\pi r^2 \rho(r) dr$ to obtain

$$\frac{dr}{dm} = \frac{1}{4\pi r^2 \rho},$$

  In operator form, the transformation between the two representations is

$$\frac{\partial}{\partial m} = \frac{1}{4\pi r^2 \rho} \frac{\partial}{\partial r}.$$

- Now we use this to convert

$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho \qquad \text{(Eulerian form)}$$

  to Lagrangian coordinates, giving

$$\frac{dP}{dm} = -\frac{Gm}{4\pi r^4} \qquad \text{(Lagrangian form).}$$

Table 4.1: Equations of hydrostatics

Eulerian coordinates (r,t)	Lagrangian coordinates (m,t)
$\dfrac{dm}{dr} = 4\pi r^2 \rho$	$\dfrac{dr}{dm} = \dfrac{1}{4\pi r^2 \rho}$
$\dfrac{dP}{dr} = -\dfrac{Gm\rho}{r^2}$	$\dfrac{dP}{dm} = -\dfrac{Gm}{4\pi r^4}$

Table 4.1 summarizes the equations of spherical hydrostatics in Eulerian and Lagrangian form.

4.3.2 Contrasting Lagrangian and Eulerian Descriptions

Eulerian or Lagrangian representations can each appear more natural in particular contexts.

- Our *observational mindset is often Eulerian*:
  - We tend to think of monitoring a river by placing a measuring device at a fixed point on the river.
  - It is less common to imagine measuring devices floating down the river with given packets of water (a Lagrangian point of view).
- The *laws of physics are often formulated in a Lagrangian way*:
  - For collision of billiard balls, we normally imagine following each ball.
  - It is less common to imagine staking out points on the table and asking how balls move past those fixed points (an Eulerian point of view).

- The Lagrangian point of view is often more simply tied to the underlying physical laws.
- Thus the Lagrangian formulation is often preferred when there are clear symmetries and conservation laws.

Example: Imagine a spherical star that is neither gaining nor losing mass, but is pulsating radially in size.

- The radial distance to the surface (Eulerian coordinate) is changing with time.
- mass contained within the outermost radius (Lagrangian coordinate) is constant in time.

- On the other hand, if
  - spherical symmetry is broken and
  - there is convective and turbulent motion of the fluid,

the Eulerian description is often simpler than the Lagrangian description.

4.4 Dynamical Timescales

A particularly important concept in astrophysics is that of a *dynamical timescale*.

A dynamical timescale sets the order of magnitude for the time required for a system to respond to a perturbation.

- The dynamical response of stars to perturbations of their hydrostatic equilibrium is of obvious significance in understanding stars and their evolution.
- Consider the *free-fall timescale* t_{ff}

$$t_{\text{ff}} \simeq \sqrt{\frac{1}{G\bar{\rho}}} \simeq \sqrt{\frac{R}{g}} \qquad \bar{\rho} = \frac{M}{\frac{4}{3}\pi R^3} \qquad g = \frac{GM}{R^2}$$

  where M is the mass, R is the radius, $\bar{\rho}$ is the average density, and g is the gravitational acceleration.

- This defines a *timescale for gravitational collapse of a uniform-density sphere* if it suddenly lost all pressure support.

- We may introduce a *second dynamical timescale* by considering the opposite extreme: *if gravity were taken away, how fast would the star expand* by virtue of its pressure?
- This timescale can depend only on R, $\bar{\rho}$, and $\bar{P}$, and the only combination of these quantities having time units is

$$t_{\rm exp} \simeq R\sqrt{\frac{\bar{\rho}}{\bar{P}}} \simeq \frac{R}{\bar{v}_{\rm s}},$$

where $v_{\rm s}$ is the average speed of sound.

This timescale has a simple interpretation:

1. $(\rho/P)^{1/2}$ is approximately the inverse of the mean sound speed $\bar{v}_{\rm s}$ for the medium.
2. This implies that $t_{\rm exp}$ is approximately the *time for a sound wave to travel from the center to the surface* of the star.

Makes sense: pressure waves propagate on that timescale.

- Hydrostatic equilibrium will be precarious unless these dynamical timescales are comparable; therefore, we define a *hydrodynamical timescale* through

$$\tau_{\rm hydro} \simeq t_{\rm exp} \simeq t_{\rm ff} \simeq \sqrt{\frac{1}{G\bar{\rho}}}.$$

Table 4.2: Hydrodynamical timescales

Object	$\sim M/M_\odot$	$\sim R/R_\odot$	$\bar{\rho}/\rho_\odot$	$\tau_{\rm hydro}$
Red Giant	1	100	10^{-6}	36 days
Sun	1	1	1	55 minutes
White Dwarf	1	1/50	10^5	9 seconds

Example: For the Sun $\bar{\rho} = 1.4\,\mathrm{g\,cm^{-3}}$ and

$$\tau_{\rm hydro} \simeq \sqrt{\frac{1}{G\bar{\rho}}} \simeq 55\,\mathrm{minutes}.$$

- If hydrostatic equilibrium were not satisfied we would expect to see changes in a matter of hours.
- But the fossil record indicates that the Sun has been extremely stable for billions of years.
- We conclude that *the Sun is in very good hydrostatic equilibrium*.

In Table 4.2 we illustrate the hydrodynamical timescale for several kinds of stars calculated using this formula.

4.5 Virial Theorem

Stars have at their disposal *two large sources of energy*:

1. *Gravitational energy,* which can be released by contraction.
2. *Internal energy,* which can be produced both by contraction and by fusion and other internal processes.

We now derive an important *relationship between internal and gravitational energy* for objects in hydrostatic equilibrium called the *virial theorem.*

Multiply both sides of the *Lagrangian hydrostatic equation*

$$\frac{dP}{dm} = -\frac{Gm}{4\pi r^4}.$$

by $4\pi r^3$ and integrate over dm from 0 to $M \equiv m(R)$ to give

$$\begin{aligned}
\int_0^M \frac{Gm}{r}\,dm &= \underbrace{-4\pi \int_0^M r^3 \frac{\partial P}{\partial m}\,dm}_{\text{Integrate by parts}} \\
&= \underbrace{-\,4\pi r^3 P\Big|_{m=0}^{m=M}}_{\text{identically zero}} + 12\pi \int_0^M r^2 P \frac{\partial r}{\partial m}\,dm \\
&= 12\pi \int_0^M r^2 P \frac{1}{4\pi r^2 \rho}\,dm = \int_0^M \frac{3P}{\rho}\,dm,
\end{aligned}$$

- ρ, r, and P are functions of independent variable m.
- An *integration by parts* was used to obtain line 2:

$$\int u dv = uv - \int v du$$

$$u = 4\pi r^3 \qquad v = P \qquad du = 12\pi r^2 dr \qquad dv = dP$$

- In the first term of line 2

  1. r vanishes when $m = 0$ (center of star), and
  2. P vanishes when $m = M$ (surface of star).

  Thus this term is identically zero.

- $\dfrac{dr}{dm} = \dfrac{1}{4\pi r^2 \rho}$ was used in going from line 2 to line 3.

The equation just obtained,

$$\int_0^M \frac{Gm}{r}\,dm = \int_0^M \frac{3P}{\rho}\,dm,$$

has a *simple interpretation*. First consider the right side:

- $P/\rho = kT/\mu$ for an ideal monatomic gas
- Thus the right side is *twice the internal energy* U because

$$\int_0^M \frac{3P}{\rho}\,dm = \frac{3kT}{\mu}\int_0^M dm = \frac{3MkT}{\mu}$$

$$= 3\underbrace{\left(\frac{M}{\mu}\right)}_{N} kT = 3NkT = 2U,$$

since for an ideal monatomic gas $U = \frac{3}{2}NkT$.

Hence we have for the right side

$$\int_0^M \frac{Gm}{r}\,dm = \int_0^M \frac{3P}{\rho}\,dm \quad \longrightarrow \quad \int_0^M \frac{Gm}{r}\,dm = 2U.$$

The integral on the left side may be interpreted by asking the question

> What is the *total gravitational energy released* in forming a star?

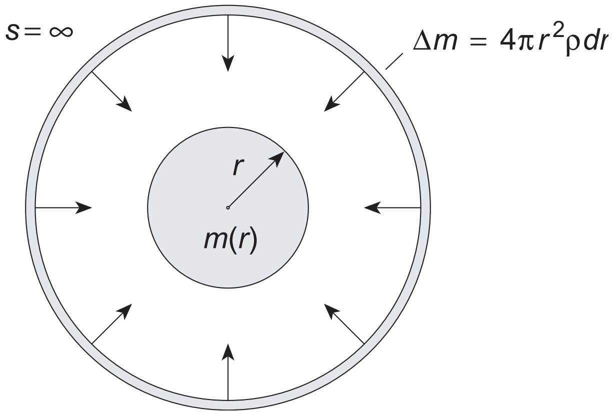


Figure 4.2: Gravitational assembly of a star by the accretion of concentric shells, each of mass $\Delta m = 4\pi r^2 \rho dr$.

Consider Fig. 4.2, where a shell of mass Δm falls from infinity onto the surface of a spherical mass of radius r and enclosed mass $m(r)$. The gravitational energy released is

$$
\begin{aligned}
d\Omega &= \int_\infty^r F_{\rm g}\, ds = \int_\infty^r g(s) \Delta m\, ds \\
&= \int_\infty^r \underbrace{\frac{Gm(r)}{s^2}}_{g(s)} \underbrace{4\pi r^2 \rho dr}_{\Delta m}\, ds \\
&= -\frac{Gm(r)}{s}\bigg|_\infty^r \times 4\pi r^2 \rho dr = -4\pi r^2 \rho dr \frac{Gm(r)}{r},
\end{aligned}
$$

and the total energy released in assembling a star of radius R and mass M from such mass shells is

$$
\Omega = \int d\Omega = -4\pi \int_0^R r^2 \rho \frac{Gm(r)}{r}\, dr = -\int_0^M \frac{Gm(r)}{r}\, dm,
$$

where $dm/dr = 4\pi r^2 \rho$ was used and $M \equiv m(R)$.

Thus, for the *left side of the virial theorem* equation

$$\int_0^M \frac{Gm(r)}{r}\,dm = -\Omega \qquad \text{(Gravitational energy of star)},$$

and from the previous result for the right side of the virial theorem equation,

$$\int_0^M \frac{Gm}{r}\,dm = 2U,$$

we see that for an ideal gas

$$\int_0^M \frac{Gm}{r}\,dm = \int_0^M \frac{3P}{\rho}\,dm \quad \longrightarrow \quad 2U + \Omega = 0$$

This result is called the *Virial Theorem* (for a monatomic ideal gas):

$$2U + \Omega = 0,$$

where U is the internal energy of the star and Ω is its gravitational energy.

The *virial theorem* for an ideal, monatomic gas,

$$2U + \Omega = 0 \qquad \text{(or in the form } U = -\tfrac{1}{2}\Omega\text{)}$$

1. Establishes a general relationship between the internal energy and gravitational energy of a star *in hydrostatic equilibrium.*

2. Is of broad applicability because

   - It was derived under very general conditions.
   - It relates the two most important energy reserves for a star:
     - *gravitational energy* and
     - *internal energy*.

We shall often use the virial theorem and concepts derived from it in discussions of stellar structure and evolution.

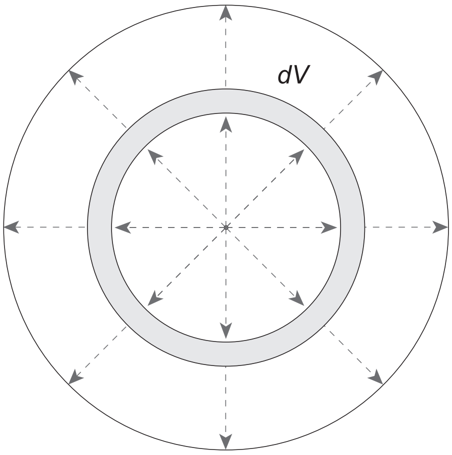


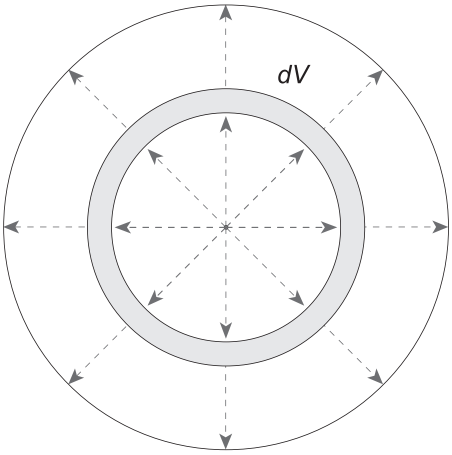
Figure 4.3: A spherical mass shell of volume dV. Dashed arrows indicate heat flow out of the star.

4.6 Thermal Equilibrium

Stars are also in approximate thermal equilibrium.

- By the *First Law*, internal energy can be changed
  - by *adding or removing heat*, or
  - by PdV *work* (expansion or contraction).
- Assume *hydrostatic equilibrium* and consider a *spherical mass shell*, as in Fig. 4.3.
- If the concentric shell is at radius r and of width dr, its volume is $dV = 4\pi r^2 dr$.
- Let's work in *Lagrangian coordinates*, with

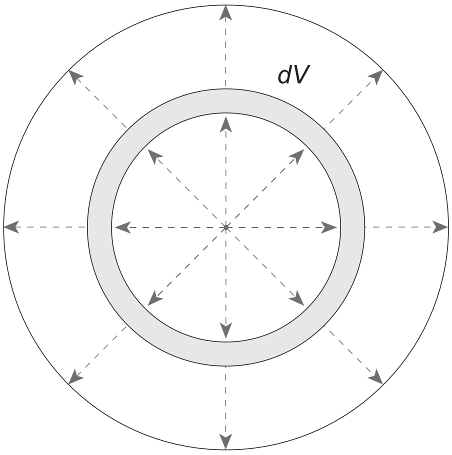
$$dm = \rho dV = 4\pi r^2 \rho dr.$$



- Let u be the *internal energy per unit mass* and
- let δf denote the change of some quantity f within the mass shell over a time t.
- The *change in heat* over a time δt is then denoted δQ and
- the *work done* in a time δt is denoted by δW.
- Then the *total change in internal energy* over a time δt is

  $$\delta(udm) = (\delta u)dm = \delta Q + \delta W,$$

  where we have used that dm is constant, by mass conservation.



As you are asked to show in *Problem 4.20* ***,

- The change in heat over a time δt is given by

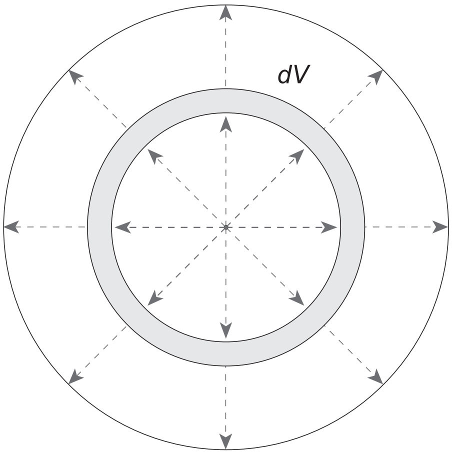
$$\delta Q = q dm \delta t - \frac{\partial L}{\partial m} dm \, \delta t,$$

and the work done in a time δt is

$$\delta W = -P \delta \left(\frac{1}{\rho} \right) dm,$$

where in these expressions

  - $L(m)$ is the luminosity associated with heat flow across the shell,
  - P is the pressure,
  - q is the rate of nuclear energy release per unit mass in the shell,
  - and where $\dfrac{dV}{dm} = \rho^{-1}$ (since $dm = \rho dV$) was used.



- Substitute

$$\delta Q = q dm \delta t - \frac{\partial L}{\partial m} dm\, \delta t \qquad \delta W = -P\delta\left(\frac{1}{\rho}\right) dm,$$

into the equation

$$\delta(udm) = (\delta u)dm = \delta Q + \delta W,$$

and take the limit $\delta t \to 0$.

- This gives a differential equation specifying the *energy balance in a mass shell* (*Problem 4.21* ***),

$$\frac{du}{dt} + P\frac{d}{dt}\left(\frac{1}{\rho}\right) = q - \frac{\partial L}{\partial m},$$

- In *thermal equilibrium* the temporal derivatives on the left side of

  $$\frac{du}{dt} + P\frac{d}{dt}\left(\frac{1}{\rho}\right) = q - \frac{\partial L}{\partial m}$$

  vanish, which implies that

  $$q = \frac{dL}{dm}.$$

- Integrate both sides over m and introduce

  $$L_0 \equiv \int_0^M q\,dm \qquad L \equiv \int_0^M \frac{dL}{dm}\,dm,$$

  where

  - L is the *total luminosity* and
  - L_0 is the *luminosity produced by nuclear reactions*.

- This leads to

  $$L_0 = L.$$

For a star in thermal and hydrostatic equilibrium, energy is radiated away at the same rate that it is produced by nuclear reactions.

4.7 Total Energy for a Star

Integrating

$$\frac{du}{dt} + P\frac{d}{dt}\left(\frac{1}{\rho}\right) = q - \frac{\partial L}{\partial m},$$

over the entire star yields

$$\int_0^M \frac{du}{dt}\,dm + \int_0^M P\frac{d}{dt}\left(\frac{1}{\rho}\right)\,dm = \int_0^M q\,dm - \int_0^M \frac{\partial L}{\partial m}\,dm.$$

The Lagrangian form of

$$\rho\,\frac{\partial^2 r}{\partial t^2} = -\frac{\partial P}{\partial r} - \frac{Gm(r)}{r^2}\rho\,.$$

is given by

$$\frac{1}{4\pi r^2}\,\frac{\partial^2 r}{\partial t^2} = -\frac{\partial P}{\partial m} - \frac{Gm(r)}{4\pi r^4}.$$

Multiplying this by $\dot{r}$ and integrating over the entire star leads to

$$\int_0^M \dot{r}\ddot{r}\,dm = -4\pi\int_0^M r^2\dot{r}\frac{\partial P}{\partial m}\,dm - \int_0^M \frac{Gm\dot{r}}{r^2}\,dm.$$

As you are asked to show in *Problem 4.22* ***,

$$\int_0^M \frac{du}{dt}\,dm + \int_0^M P\frac{d}{dt}\left(\frac{1}{\rho}\right)\,dm = \int_0^M q\,dm - \int_0^M \frac{\partial L}{\partial m}\,dm.$$

together with

$$\int_0^M \dot{r}\ddot{r}\,dm = -4\pi\int_0^M r^2\dot{r}\frac{\partial P}{\partial m}\,dm - \int_0^M \frac{Gm\dot{r}}{r^2}\,dm.$$

imply an energy-conservation equation

$$\dot{E} = \dot{U} + \dot{\Omega} + \dot{K} = L_0 - L,$$

where

- dots indicate time derivatives,
- the total energy is $E = U + K + \Omega$,
- U is the total internal energy,
- Ω is the total gravitational energy,
- The total kinetic energy is $K = \frac{1}{2}\int_0^M \dot{r}^2\,dm$,
- L is the total luminosity, and
- L_0 is the luminosity deriving from nuclear reactions.

The *total energy* of a star is

$$E = U + K + \Omega,$$

and the *equation of energy conservation* is

$$\dot{E} = \dot{U} + \dot{\Omega} + \dot{K} = L_0 - L,$$

- If the star is in *thermal equilibrium* $\dot{E} = 0$ and
- if it is in *hydrostatic equilibrium* $K = 0$.

In the limit of *hydrostatic and thermal equilibrium*, properties of stars are governed by the *virial theorem* relating U to Ω.

4.8 Stability and Heat Capacity

We have argued above that

- stars are in a *hydrostatic equilibrium* that balances gravitational forces against pressure-differential forces, and
- a *thermal equilibrium* that balances energy production against energy emission.

But how *stable* is that equilibrium?

- A ball at the bottom of a deep valley and
- a ball balanced on a knife edge

are both in equilibrium, but they have *very different stabilities*.

- Are stars *in a deep valley*, or
- are they *balanced on a knife edge*?

As we shall see, the answer

- depends very much on the *equation of state*, and
- is the source of both
  - the *remarkable stability of main sequence stars*, and
  - some of the most *violent explosions* observed in our Universe.

We will address a number of instabilities in later chapters; here we illustrate for *thermal instability*.

4.8.1 Temperature Response to Energy Fluctuations

Consider a star with an ideal gas plus radiation equation of state

$$P = P_g + P_r = nkT + \frac{1}{3}aT^4 = \frac{2}{3}u_g + \tfrac{1}{3}u_r,$$

where we assume for the ideal gas an internal energy density

$$u_g = \frac{3}{2}nkT = \frac{3}{2}P_g$$

and for the radiation

$$u_r = aT^4 = 3P_r.$$

Then the gravitational energy is

$$\begin{aligned}\Omega &= -\int_0^M \frac{3P}{\rho}\,dm \\ &= -2\int_0^M \frac{u_g}{\rho}\,dm - \int_0^M \frac{u_r}{\rho}\,dm \\ &= -2U_g - U_r,\end{aligned}$$

since the total internal energies are given by

$$U_g = 4\pi\int_0^R u_g r^2\,dr = \int_0^M \frac{u_g}{\rho}\,dm,$$

$$U_r = 4\pi\int_0^R u_r r^2\,dr = \int_0^M \frac{u_r}{\rho}\,dm.$$

Thus, using $\Omega = -2U_g - U_r$ the total energy is

$$E = \Omega + U_r + U_g = -U_g = -\frac{3}{2}NkT.$$

Letting

- L denote the *luminosity* of the star and
- L_0 the *energy generation rate*,

their difference may be written as

$$L_0 - L = \frac{dE}{dt} = -\frac{3}{2}Nk\frac{dT}{dt},$$

where the derivative was evaluated using

$$E = \Omega + U_{\rm r} + U_{\rm g} = -U_{\rm g} = -\tfrac{3}{2}NkT.$$

At thermal equilibrium $L_0 - L = 0$.

At thermal equilibrium

$$L_0 - L = -\frac{3}{2}Nk\frac{dT}{dt} = 0.$$

Now suppose a small fluctuation away from equilibrium occurs such that

$$L_0 - L = \delta L = -\frac{3}{2}Nk\frac{dT}{dt}.$$

Solving for dT/dt gives

$$\frac{dT}{dt} = -\frac{2}{3}\frac{\delta L}{Nk},$$

which governs how the temperature will respond to *small fluctuations in energy production* (or energy transport).

The response of temperature T to an energy fluctuation δL is

$$\frac{dT}{dt} = -\frac{2}{3}\frac{\delta L}{Nk}.$$

Now consider two situations for $\delta L \equiv L_0 - L$:

1. If $\delta L > 0$, rate of energy generation *exceeds luminosity*:

   $$\delta L > 0 \quad \rightarrow \quad L_0 > L \quad \rightarrow \quad dT/dt < 0.$$

   Thus the response to an *increase in energy generation* is

   - a *decrease in temperature*,
   - which tends to *decrease the energy generation rate*.

2. If $\delta L < 0$, energy generation rate is *less than luminosity*:

   $$\delta L < 0 \quad \rightarrow \quad L_0 < L \quad \rightarrow \quad dT/dt > 0.$$

   Thus the response to a *decrease in energy generation* is

   - an *increase in temperature*, which causes
   - an *increase in the rate of energy generation*.

These responses are the *essence of a stable system*:

- An imbalance causes an *automatic restorative action* that re-establishes the balance.
- Yet this essential feature of normal stars is quite *counter-intuitive*!

As a gas cloud contracts to form a star, gravitational energy $\Delta\Omega$ is released.

- The slowly collapsing protostar goes through a sequence of stages that are *nearly in hydrostatic equilibrium*.
- *The virial theorem must be satisfied* for hydrostatic equilibrium to hold.
- Thus, as a newly-forming star contracts the *virial theorem must be satisfied approximately*, which requires that

- The *thermal energy* must change by

$$\Delta U \simeq -\tfrac{1}{2}\Delta\Omega,$$

- and the *excess energy must be radiated away* before the star can contract further.

Hence, gravitational contraction has three consequences:

1. The star *heats up*,
2. Some energy is *radiated into space*,
3. The *star's total energy decreases* and *it becomes more bound*.

Stated concisely: the star *"heats up while it cools down"*.

How can a star *"heat up while it cools down"?*

Answer: *GRAVITY.*

The *virial theorem* is

$$2U = -\Omega.$$

Identifying U as the kinetic energy and Ω as the potential energy, an alternative statement is

$$E_{\mathrm{kin}} = -\tfrac{1}{2}E_{\mathrm{pot}},$$

and if $E = E_{\mathrm{kin}} + E_{\mathrm{pot}}$ is the total energy,

$$E_{\mathrm{kin}} = -E.$$

But this implies that

- Adding energy *decreases the kinetic energy.*
- Removing energy *increases the kinetic energy.*

Thus, identifying *average kinetic energy* ↔ *temperature*,

- *adding energy decreases* T;
- *removing energy increases* T.

Stars have *negative heat capacity* since *gravity is long-ranged*.

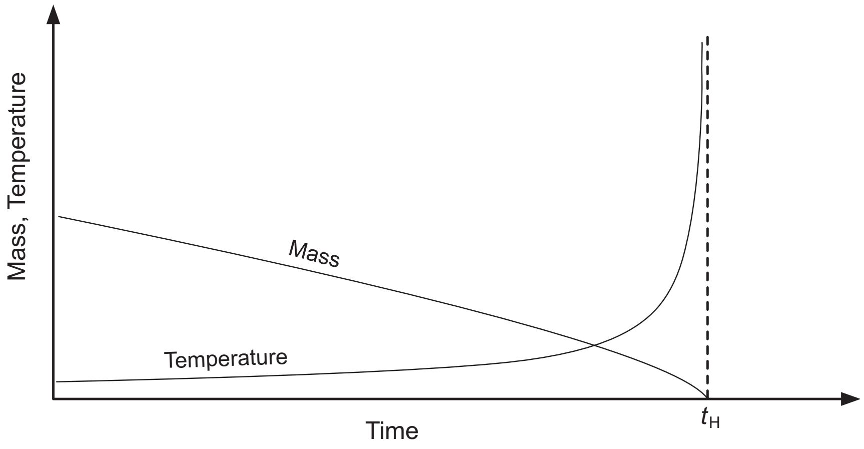
- Counterintuitive: We find almost all local objects to have *positive heat capacities*.
- But local objects *aren't bound by gravity!*

Another Example (from Astronomy 421): Because of quantum *Hawking radiation,* black holes have a temperature

$$T = \frac{\hbar c^3}{8\pi kGM},$$

where M is the mass, k is Boltzmann's constant, $\hbar$ is Planck's constant divided by 2π, and G is the gravitational constant. As Hawking radiation is emitted

1. the black hole *loses mass (energy)*, and
2. its *temperature rises:* $T \to \infty$ as $M \to 0$.



Like a star, a black hole (the ultimate gravitating system!) becomes hotter as it loses energy: it exhibits a *negative heat capacity.*

4.9 Kelvin–Helmholtz Timescale for the Sun

Returning to our contracting protostar, if approximate hydrostatic equilibrium is to be maintained, the *virial theorem* requires that the *thermal energy* must change by

$$\Delta U \simeq -\frac{1}{2}\Delta\Omega.$$

1. Thus, at each infinitesimal step of the contraction

   the *star must wait* until half of the released gravitational energy is radiated

   before it can continue to contract.

2. This implies that there is a *timescale for contraction* in near hydrostatic equilibrium that is set by the *time required to radiate the excess energy*.

This contraction timescale is called the *Kelvin–Helmholtz timescale* or the *thermal adjustment timescale*.

We may estimate the *Kelvin–Helmholtz timescale* by assuming constant density ρ and a corresponding mass contained within the radius r

$$m(r) = \frac{4}{3}\pi r^3 \rho,$$

during the collapse. Then the *gravitational energy released* in collapsing down to a star of radius R is

$$\begin{aligned}\Omega &= -\int_0^R 4\pi r^2 \rho \frac{Gm}{r}\, dr \qquad (\text{ substitute } m = \tfrac{4}{3}\pi r^3 \rho) \\ &= -\frac{16}{15}\pi^2 \rho^2 G R^5 \qquad (\text{ substitute } \rho = \frac{3M}{4\pi R^3}) \\ &= -\frac{3}{5}\frac{GM^2}{R},\end{aligned}$$

where the total mass is

$$M = \frac{4}{3}\pi R^3 \rho .$$

Taking $M = M_\odot$ and $R = R_\odot$, we find that $\Omega_\odot = 2.3 \times 10^{48}$ erg of gravitational energy was released in forming the Sun. By the *virial theorem*, half of this must be radiated while the Sun contracts:

$$E_{\text{rad}}^{\odot} = \frac{1}{2}\Omega_\odot \simeq \frac{GM_\odot^2}{R_\odot} \simeq 10^{48}\,\text{erg}.$$

The *Kelvin–Helmholtz timescale* t_{KH} is the characteristic time to radiate this energy.

We may make a rough estimate of the *Kelvin–Helmholtz timescale for the Sun* by assuming that it has had its present luminosity of $L_\odot = 4 \times 10^{33}\,\mathrm{erg\,s^{-1}}$ for its entire life. Then

$$t_{\rm KH} \simeq \frac{E_{\rm rad}^{\odot}}{L_\odot} \simeq \frac{GM_\odot^2/R_\odot}{L_\odot} \simeq 10^7 \text{ years},$$

We conclude that the Sun *contracted to the main sequence* on a Kelvin–Helmholtz timescale of about *10 million years.*

Generally, we shall define a *Kelvin–Helmholtz timescale for a star* by

$$t_{\rm KH} = \frac{\Omega}{L} \simeq \frac{GM^2/R}{L},$$

where R is the radius, M the mass, and L the luminosity.

4.10 Kelvin–Helmholtz Timescale for Other Stars

The *Kelvin–Helmholtz timescale for other stars* may be related to that of the Sun by scaling. Since generally

$$t_{\mathrm{KH}} = \frac{\Omega}{L} \simeq \frac{GM^2}{LR},$$

the ratio of the Kelvin–Helmholtz timescale for some star relative to that of the Sun is given by

$$\frac{t_{\mathrm{KH}}}{t_{\mathrm{KH}}^{\odot}} = \left(\frac{R_\odot}{R}\right)\left(\frac{L_\odot}{L}\right)\left(\frac{M}{M_\odot}\right)^2,$$

where a good estimate is $t_{\mathrm{KH}}^{\odot} = 3 \times 10^7$ yr.

Example: From Table 2.2 in the book, an *A0 main sequence star* like Sirius A has

$$R = 2.5 R_\odot \quad M = 3.2 M_\odot \quad L = 79.4 L_\odot .$$

Inserting these values in the above equation, the *Kelvin–Helmholtz timescale* for an A0 star is

$$t_{\mathrm{KH}} \sim 0.055\, t_{\mathrm{KH}}^{\odot} \sim 1.55 \times 10^6 \text{ yr}.$$

More massive stars evolve more rapidly through all phases of their lives, including periods of gravitational contraction.

Chapter 5

Thermonuclear Reactions in Stars

Stars have three primary sources of energy:

1. *heat left over* from earlier processes,
2. *gravitational energy*, and
3. energy released by *thermonuclear reactions*.

We shall see that

- Gravitational energy is important in star birth, star death, and various transitional stages.
- White dwarfs shine because of heat left over from earlier energy generation.
- But the *virial theorem* indicates that gravity and left-over heat can power Sun only on a 10^7 *year (Kelvin–Helmholtz) timescale*.

> *Thermonuclear reactions are the only viable long-term stellar energy source.*

5.1 Nuclear Energy Sources

The measured *luminosity of the Sun* is

$$L_\odot \simeq 3.8 \times 10^{33}\,\mathrm{erg\,s^{-1}}$$

and that of the *most luminous stars* is about $10^6 L_\odot$.

- From the *Einstein relation* $E = mc^2$,

  $$\Delta m = \frac{\Delta E}{c^2},$$

  and the *rate of mass conversion to energy* required to sustain the Sun's luminosity is

  $$\Delta m = \frac{1}{c^2}\Delta E \to \frac{\Delta m}{\Delta t} = \frac{1}{c^2}\frac{\Delta E}{\Delta t} = \frac{L_\odot}{c^2} = 4.2 \times 10^{12}\,\mathrm{g\,s^{-1}}.$$

- The most luminous stars require conversion rates a million times larger.

Let us now discuss how nuclear reactions in stars can account for mass-to-energy conversion on this scale.

5.1.1 The Curve of Binding Energy

The *binding energy* for a nucleus of atomic number Z and neutron number N is

$$B(Z,N) \equiv [\underbrace{Zm_{\rm p} + Nm_{\rm n}}_{\text{free nucleons}} - \underbrace{m(Z,N)}_{\text{bound system}}]c^2,$$

where

- $m(Z,N)$ is the mass of the nucleus
- $m_{\rm p}$ is the mass of a proton
- $m_{\rm n}$ is the mass of a neutron.

The binding energy may be interpreted either as

- the *energy released in assembling a nucleus from its constituent nucleons*, or
- the *energy required to break a nucleus apart into its constituents*.

The more relevant quantity is often the *binding energy per nucleon,* $B(Z,N)/A$, where $A = Z + N$ is the atomic mass number.

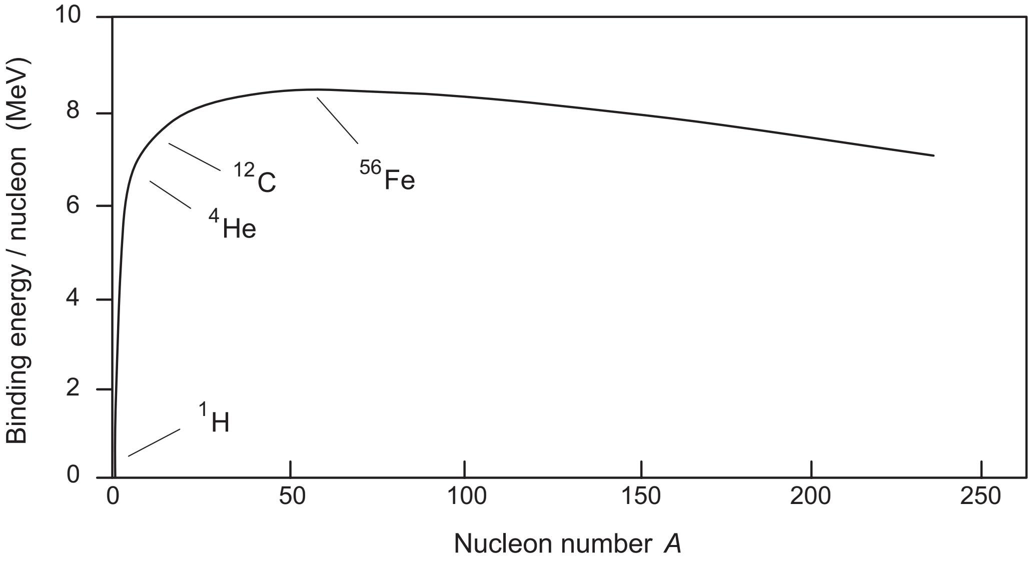


Figure 5.1: The curve of binding energy. Only the average behavior is shown; local fluctuations have been suppressed, as has the isotopic dependence on (Z,N) for a given A.

- The average behavior of binding energy per nucleon as a function of the atomic mass number A is shown in Fig. 5.1.
- The general behavior of the curve of binding energy may be understood from simple nuclear physics considerations.
- These considerations are elaborated in Chapter 5 of the book but we won't go over them here.

5.1.2 Masses and Mass Excesses

It is convenient to define the *mass excess,* $\Delta(A,Z)$, through

$$\Delta(A,Z) \equiv (m(A,Z) - A)\, M_u c^2,$$

- $m(A,Z)$ is measured in atomic mass units (amu),
- $A = Z + N$ is the atomic mass number, and
- the atomic mass unit M_u (which is defined to be $\frac{1}{12}$ the mass of a ^{12}C atom) is given by

  $$M_u = \frac{1}{N_{\mathrm{A}}} = 1.660420 \times 10^{-24}\,\mathrm{g} = 931.478\,\mathrm{MeV}/c^2,$$

  with $N_{\mathrm{A}} = 6.02 \times 10^{23}\,\mathrm{mole}^{-1}$ (Avogadro's constant).

The mass excess is useful because

- The number of nucleons (neutrons + protons) is constant in low-energy nuclear reactions,
- so atomic mass numbers cancel on both sides of equations.
- Thus, sums and differences of masses (large numbers) may be replaced by the corresponding sums and differences of mass excesses (small numbers).

Modern computers don't care. But when a theory of nuclear masses was developed in the 1930s and 1940s, *computers were people working by hand*.

Example: Using the definition of the mass excess, the binding energy equation

$$B(Z,N) \equiv [Zm_{\rm p} + Nm_{\rm n} - m(Z,N)]c^2,$$

may be rewritten as

$$\begin{aligned} B(Z,N) &= [Zm_{\rm p} + Nm_{\rm n} - m(Z,N)]c^2 \\ &= [Zm_{\rm p} + (A-Z)m_{\rm n} - m(Z,N)]c^2 \\ &= [Z\Delta_{\rm p} + (A-Z)\Delta_{\rm n} - \Delta(A,Z)]c^2 \\ &= [Z\Delta_{\rm p} + (A-Z)\Delta_{\rm n} - \Delta(A,Z)] \times 931.478\,\text{MeV}, \end{aligned}$$

where we have

- abbreviated the mass excess of the neutron and proton by $\Delta(1,0) \equiv \Delta_{\rm n}$ and $\Delta(1,1) \equiv \Delta_{\rm p}$, respectively,
- and in the last line *units of amu are assumed*.

In many mass tables, the mass excesses rather than the masses themselves are tabulated.

Example: Let's calculate the binding energy of ^{4}He. The relevant mass excesses are

$$\Delta_{\rm p} = 7.289\,\text{MeV} \quad \Delta_{\rm n} = 8.071\,\text{MeV} \quad \Delta(4,2) = 2.425\,\text{MeV}$$

and the binding energy of ^{4}He is then

$$\begin{aligned} B(Z,N) &= [Z\Delta_{\rm p} + (A-Z)\Delta_{\rm n} - \Delta(A,Z)]c^2 \\ &= 2\times 7.289 + 2\times 8.071 - 2.425 = 28.3\,\text{MeV}. \end{aligned}$$

Thus more that 28 MeV of energy is required to separate ^{4}He into free neutrons and protons.

5.1.3 Q-Values

The *Q-value for a reaction* is

- the total mass of the reactants minus the total mass of the products,
- which is equivalent to the corresponding difference in mass excesses:

$$\begin{aligned} Q &= \text{Mass of reactants} - \text{Mass of products} \\ &= \text{Mass excess of reactants} - \text{Mass excess of products.} \end{aligned}$$

It is common to specify the Q-value in *energy units rather than mass units* (by multiplying masses by c^2).

Example: For the nuclear reaction

$$^{2}\mathrm{H} + {}^{12}\mathrm{C} \to {}^{1}\mathrm{H} + {}^{13}\mathrm{C},$$

the tabulated mass excesses are

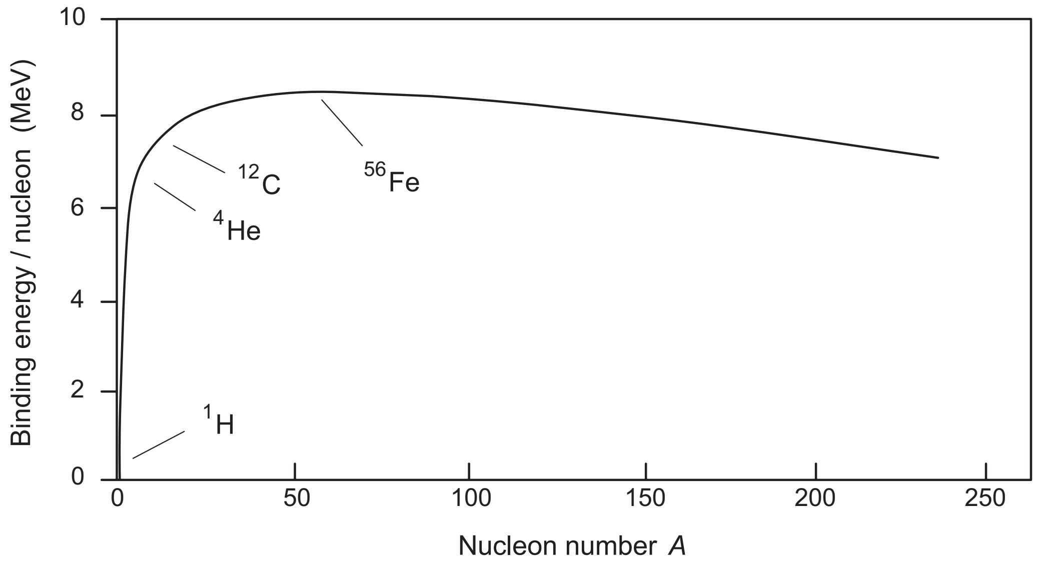
$$\Delta(^{2}\mathrm{H}) = 13.136\,\mathrm{MeV} \qquad \Delta(^{12}\mathrm{C}) = 0\,\mathrm{MeV}$$

$$\Delta(^{1}\mathrm{H}) = 7.289\,\mathrm{MeV} \qquad \Delta(^{13}\mathrm{C}) = 3.1246\,\mathrm{MeV}.$$

The Q-value for this reaction is then

$$Q = \Delta(^{2}\mathrm{H}) + \Delta(^{12}\mathrm{C}) - \Delta(^{1}\mathrm{H}) - \Delta(^{13}\mathrm{C}) = +2.72\,\mathrm{MeV}.$$

- The *positive value of Q* indicates that this is an *exothermic reaction:*
  - 2.72 MeV is *liberated from binding energy* in the reaction.
  - This appears as *kinetic energy* or *internal excitation* of the products.
- Conversely, a *negative value of Q*
  - indicates an *endothermic reaction.*
  - This means that *additional energy must be supplied* to make the reaction viable.



5.1.4 Efficiency of Hydrogen Fusion

Examination of the curve of binding energy suggests two potential nuclear sources of energy:

- *Fission* of heavier elements into lighter elements.
- *Fusion* of lighter elements into heavier ones.

Since stars are composed mostly of hydrogen and helium,

- Their energy source must be *fusion of lighter elements*.
- *Coulomb repulsion* between charged nuclei will inhibit fusion, so *hydrogen* ($Z=1$) *will be easier to fuse than helium* ($Z=2$).
- In particular, main sequence stars are powered by *thermonuclear processes that convert four* ^{1}H *into* ^{4}He.

The total *rest mass energy in one gram* of material is

$$E = mc^2 = 9 \times 10^{20}\,\text{erg},$$

and the energy released in the conversion of one gram of hydrogen into ^{4}He is

$$\Delta E\,(\text{fusion H}\rightarrow{}^4\text{He}) = 6.3 \times 10^{18}\,\text{erg g}^{-1}.$$

Therefore,

- *less than 1% of the initial rest mass* is converted into energy in the stellar fusion of hydrogen into helium:

$$\frac{\Delta E\,(\text{fusion H}\rightarrow{}^4\text{He})}{\text{total rest-mass energy}} = \frac{6.3 \times 10^{18}\,\text{erg g}^{-1}}{9 \times 10^{20}\,\text{erg g}^{-1}} \simeq 0.007.$$

- We see from these considerations that hydrogen fusion is a rather *inefficient* source of energy.
- Furthermore, fusion rates in the cores of lower-mass main sequence stars are *quite small*.

The Sun's luminosity is equivalent to *several 100-watt lightbulbs per cubic meter* of the core.

- The reason that fusion is able to power stars is *not because of its intrinsic efficiency*.
- Rather it is because of the *enormous mass of stars*, which implies that they have *large reservoirs of hydrogen* fuel.

5.2 Thermonuclear Hydrogen Burning

The primary energy source of main sequence stars derives from *conversion of hydrogen into helium*.

Two sets of *thermonuclear reactions* can accomplish this:

1. the *proton–proton chain* (*PP chain*), and
2. the *CNO cycle*

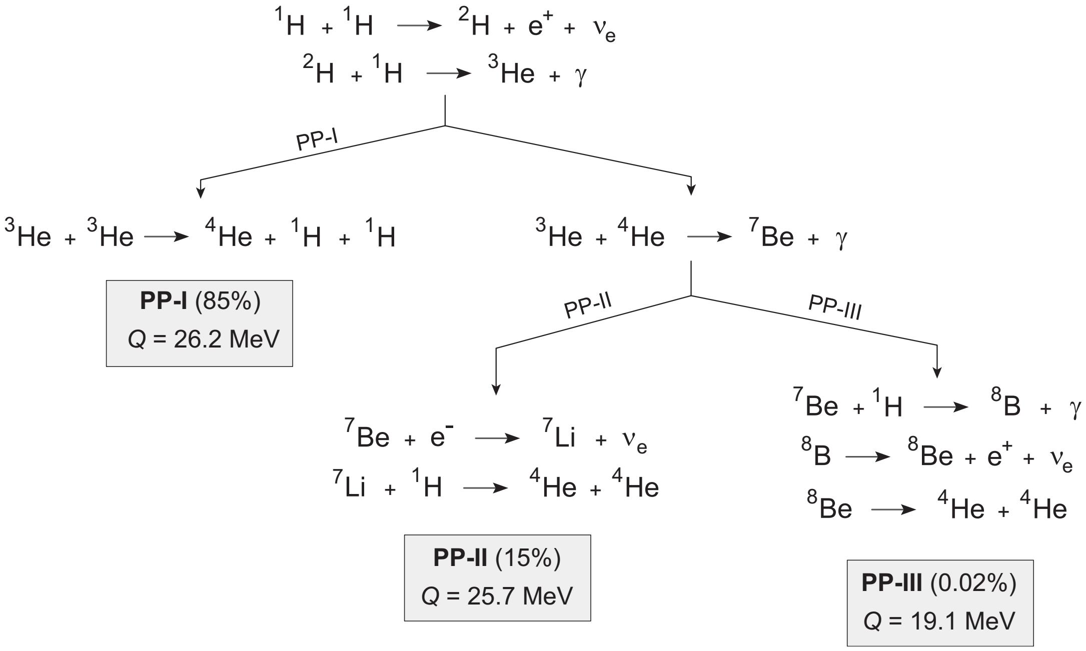
Generally it is found that

- The *proton–proton chain*
  - *produces most of the energy of the Sun* and
  - generally is *dominant in stars of a solar mass or less*.
- The *CNO cycle* quickly surpasses the proton–proton chain in energy production as soon as the mass exceeds about a solar mass.

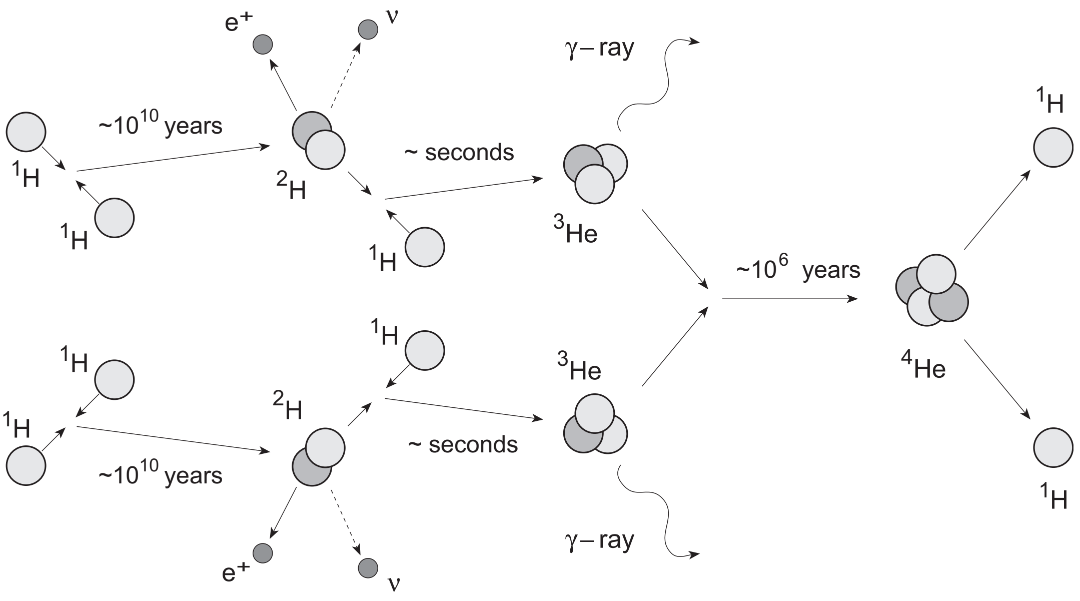
The reason for this rapid switchover is that the PP chain and the CNO cycle have strong and very different *dependence on temperature.*

5.2.1 The Proton–Proton Chains

The proton–proton (PP) chains are summarized below:


$^1H + {^1H} \longrightarrow {^2H} + e^+ + \nu_e$
$^2H + {^1H} \longrightarrow {^3He} + \gamma$
PP-I
$^3He + {^3He} \longrightarrow {^4He} + {^1H} + {^1H}$
PP-I (85%)
$Q = 26.2$ MeV
$^3He + {^4He} \longrightarrow {^7Be} + \gamma$
PP-II
PP-III
$^7Be + e^- \longrightarrow {^7Li} + \nu_e$
$^7Li + {^1H} \longrightarrow {^4He} + {^4He}$
PP-II (15%)
$Q = 25.7$ MeV
$^7Be + {^1H} \longrightarrow {^8B} + \gamma$
$^8B \longrightarrow {^8Be} + e^+ + \nu_e$
$^8Be \longrightarrow {^4He} + {^4He}$
PP-III (0.02%)
$Q = 19.1$ MeV

Cartoon of PP-I:



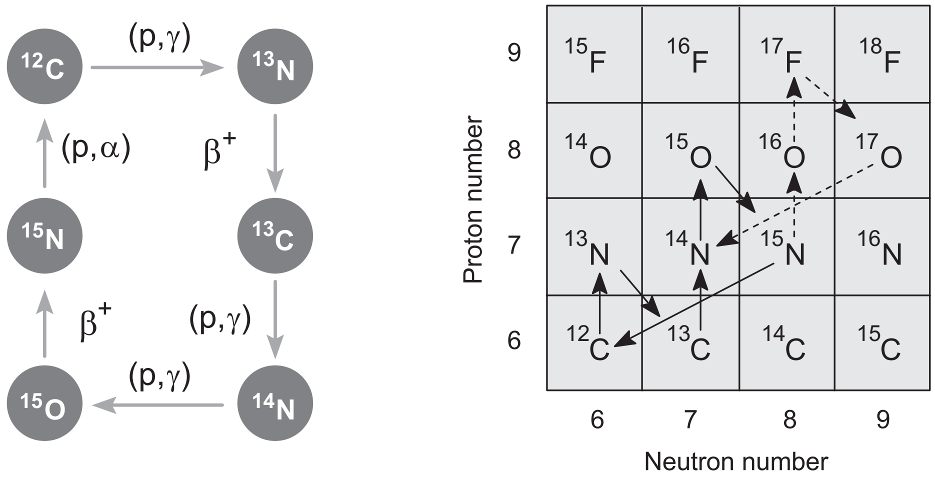


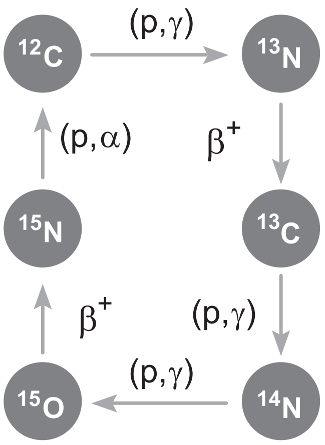
Figure 5.2: The CNO cycle. The main part of the cycle is illustrated schematically on the left side. On the right side the main part of the cycle is illustrated with solid arrows and a side branch is illustrated with dashed arrows. The notation (p,i) means a proton capture followed by emission of i; for example $^{12}\mathrm{C}(p,\gamma)^{13}\mathrm{N}$. β^+ indicates beta decay by positron emission; for example, $^{13}\mathrm{N} \to {}^{13}\mathrm{C} + e^+ + \nu_\mathrm{e}$.

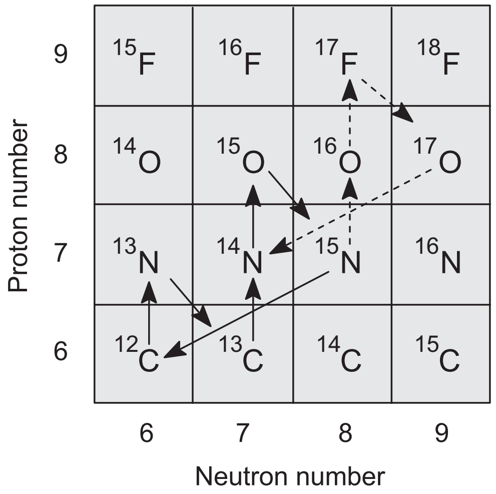
5.2.2 The CNO Cycle

The name of the carbon–nitrogen–oxygen or *CNO cycle*

- derives from the role played by isotopes of
  - carbon (C),
  - nitrogen (N), and
  - oxygen (O)

  in the corresponding sequence of reactions.
- The *CNO cycle* is summarized in Fig. 5.2 above.





CNO Catalysis

The main part of the CNO cycle is termed the *CN cycle.*

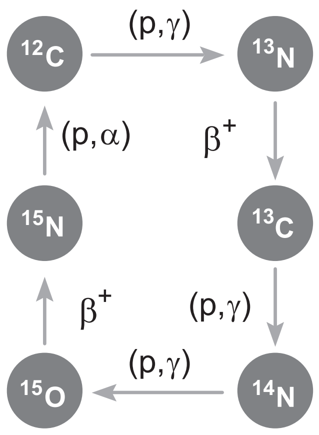
- Summing net reactants and products around the CN cycle,

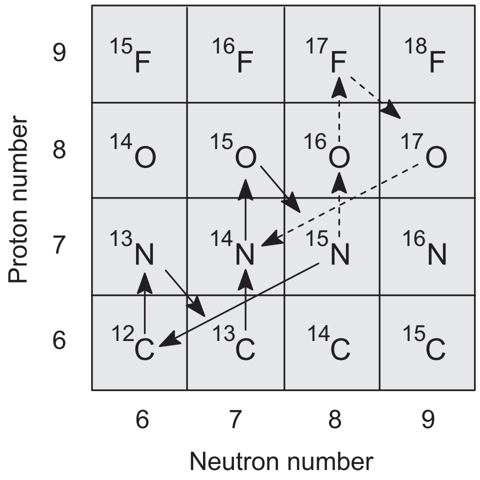
$$\underline{^{12}\mathrm{C}} + 4p \quad \longrightarrow \quad \underline{^{12}\mathrm{C}} + {}^{4}\mathrm{He} + 2\beta^{+} + 2\nu .$$

  (The γ-rays have been neglected since they *do not correspond to a conserved quantity*.)

- Therefore, ^{12}C serves as a *catalyst* for the conversion of four protons to ^{4}He.
- It is *required for the sequence to take place*, but
- it is *not consumed in the process*, because a ^{12}C is returned in the last step of the cycle.

The Q-value for the main CNO cycle is 23.8 MeV and it supplies *less than 2%* of the Sun's energy.





We have written the CNO sequence as if *the (p,γ) reaction on* ^{12}*C were the first step*,

- but CNO is a *closed cycle*.
- Hence we may consider *any step to be the initial one*.
- This implies that *any of the C, N, or O isotopes* in the cycle may be viewed as the catalyst that converts protons into helium.
- The *closed nature of the cycle* also implies that

  1. *Any mixture of these isotopes* will play the same catalytic role.
  2. If *any one of the CNO isotopes is present initially* a mixture of the others will inevitably be produced by the cycle of reactions.

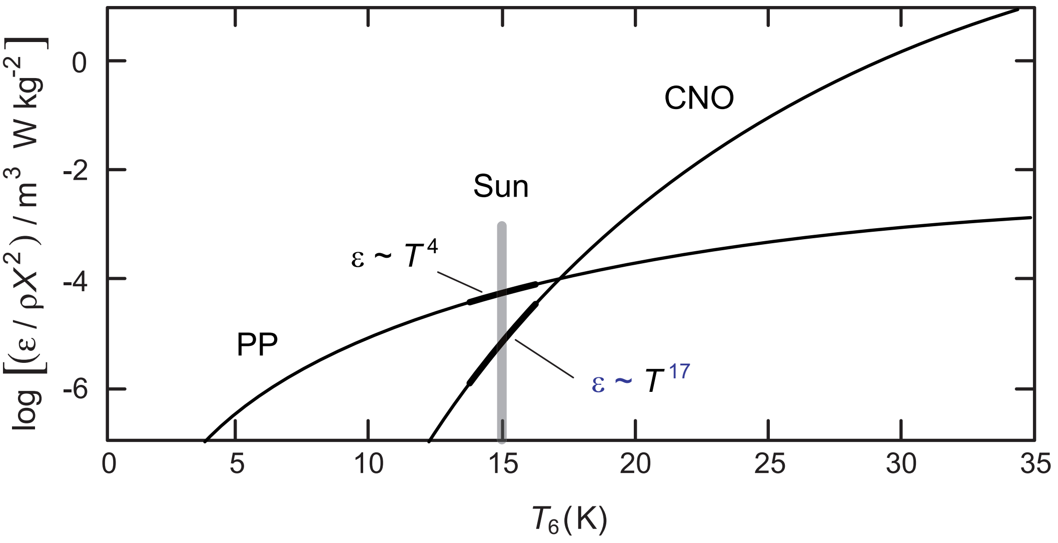


Figure 5.3: Rate of energy release in the PP chain and in the CNO cycle. T_6 denotes the temperature in units of 10^6 K.

Rates of energy release from hydrogen burning for the PP chain and CNO cycle are illustrated in Fig. 5.3.

We will see how to calculate these curves later in this chapter.

- PP chains have a strong temperature dependence $(\sim T^4)$,
- but the CNO cycle has a *even stronger dependence* $(\sim T^{17})$.
- This temperature dependence implies that
  - the *star's mass on the main sequence* is the most important factor governing the PP–CNO competition,
  - because *mass strongly influences core temperature*.

The PP cycle can occur in any star containing H, but the CNO cycle requires the presence of C, N, or O as catalysts.

- Therefore, the CNO cycle should be relatively more important in Pop I stars.
- However, this is generally of *secondary importance to temperature*, as long as some CNO isotopes are present.
- This issue is also of importance in understanding the very first generation of stars (*Pop III*).
  - *No CNO isotopes* were produced in the big bang.
  - Thus the *first generation of stars operated by the PP chain* until some of those stars could produce and distribute carbon.
- These considerations are important for the early Universe because relatively massive stars formed surprisingly early.
  1. *CNO is more efficient* in massive stars than PP because of its temperature dependence.
  2. Thus the *pace of early structure evolution*
     - depended on when the earliest stars produced CNO isotopes,
     - which then allowed succeeding generations of stars to *switch to the more efficient CNO cycle*.

5.3 Cross Sections and Reaction Rates

A quantitative analysis of energy production in stars requires the basics of *nuclear reaction theory for stellar environments*.

- Let's begin by considering the nuclear reaction

  $$\alpha + \mathrm{X} \longrightarrow \mathrm{Z}^* \longrightarrow \mathrm{Y} + \beta,$$

  where Z^* denotes an excited intermediate state called a *compound nucleus*.

- *Note:* we will often write this equation in the nuclear physics notation as $\mathrm{X}(\alpha,\beta)\mathrm{Y}$.

- A *compound nucleus* is
  - an *excited composite* that
  - *quickly decays into the final products* of the reaction.

- In the reaction

  $$\underbrace{\alpha + \mathrm{X}}_{\text{entrance channel}} \longrightarrow \mathrm{Z}^* \longrightarrow \underbrace{\mathrm{Y} + \beta}_{\text{exit channel}}$$

  - the left side $(\alpha + X)$ is called the *entrance channel* and
  - the right side $(Y + \beta)$ is called the *exit channel*

  for the reaction.

- It is common to classify nuclear reactions according to the *number of (nuclear) species in the entrance channel*; thus

$$\alpha + \mathrm{X} \longrightarrow \mathrm{Z}^* \longrightarrow \mathrm{Y} + \beta,$$

  is a *2-body reaction.*
- We shall often use 2-body reactions to illustrate but
- *1-body reactions* of the form $A \to B + C$ and
- *3-body reactions* of the form $A + B + C \to D$

also play a role in stellar energy production.

Let us first consider a *laboratory experiment* where

- The reaction is *initiated by a beam of projectiles* α directed onto a *target containing nuclei* X.
- The *cross section* $\sigma_{\alpha\beta}(v)$ is defined by

$$\sigma_{\alpha\beta}(v) \equiv \frac{\rho_{\alpha\beta}}{F(v)}$$
$$= \left(\frac{\text{reactions per unit time per target nucleus}}{\text{incident flux of projectiles}} \right).$$

  It is a function of the velocity v, and has *units of area*.

> A common *unit of cross section* is the *barn (b)*, which is a *cross section of* $10^{-24}\,cm^2$.

- The *incident particle flux* $F(v)$ is given by

$$F(v) = n_\alpha v,$$

  where

  - n_α is the *number density of projectiles* α in the beam and
  - v is their *velocity*.

- The number of reactions per unit time (*reaction rate*) per target nucleus $\rho_{\alpha\beta}$ is

  $$\rho_{\alpha\beta} = F_v \sigma_{\alpha\beta} = n_\alpha v \sigma_{\alpha\beta},$$

  and the *total reaction rate per unit volume* $r_{\alpha\beta}(v)$ results from

  - multiplying $\rho_{\alpha\beta}$ by the number density n_X of target nuclei X:

    $$r_{\alpha\beta}(v) = \rho_{\alpha\beta} n_X = n_\alpha n_\mathrm{X} v \sigma_{\alpha\beta}(v)(1+\delta_{\alpha\mathrm{X}})^{-1},$$

  - and has units of $\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$ in the CGS system.

- The factor involving the Kroenecker δ_{ab} in

$$r_{\alpha\beta}(v) = \rho_{\alpha\beta} n_X = n_\alpha n_X v \sigma_{\alpha\beta}(v)(1+\delta_{\alpha\mathrm{x}})^{-1},$$

(where δ_{ab} is one if $a = b$ and zero if $a \neq b$) is introduced to *prevent overcounting when the colliding particles are identical*.

1. The product $n_\alpha n_X v \sigma_{\alpha\beta}(v)$ is the rate per unit volume for the 2-body reaction
2. $n_\alpha n_X$ is the number of unique particle pairs (α, X) contained in the unit volume.
3. But for the collision of identical particles $(\alpha = X)$, the number of independent particle pairs (α, α) is not N_α^2 but $\frac{1}{2}N_\alpha^2$.
4. Therefore, for identical particles the rate expression must be multiplied by a factor of $1/(1+\delta_{\alpha X}) = \frac{1}{2}$ to avoid double counting.
5. More generally, for N identical particles a factor of $1/N!$ is required to prevent double counting.

- We will not display the Kroenecker-δ factors unless they are essential to the discussion.

- Normally we will work in the *center of mass (CM) coordinate system.*
- Thus velocities, energies, momenta, and cross sections will be *center of mass quantities*, with

$$E \equiv E_{\mathrm{CM}} = \left(\frac{m_{\mathrm{X}}}{m_\alpha + m_{\mathrm{X}}}\right) E_{\mathrm{lab}},$$

$$v \equiv v_{\mathrm{CM}} = \sqrt{\frac{2E}{\mu}},$$

$$\mu \equiv \frac{m_\alpha m_{\mathrm{X}}}{m_\alpha + m_{\mathrm{X}}} \qquad \text{(reduced mass)},$$

unless otherwise noted.

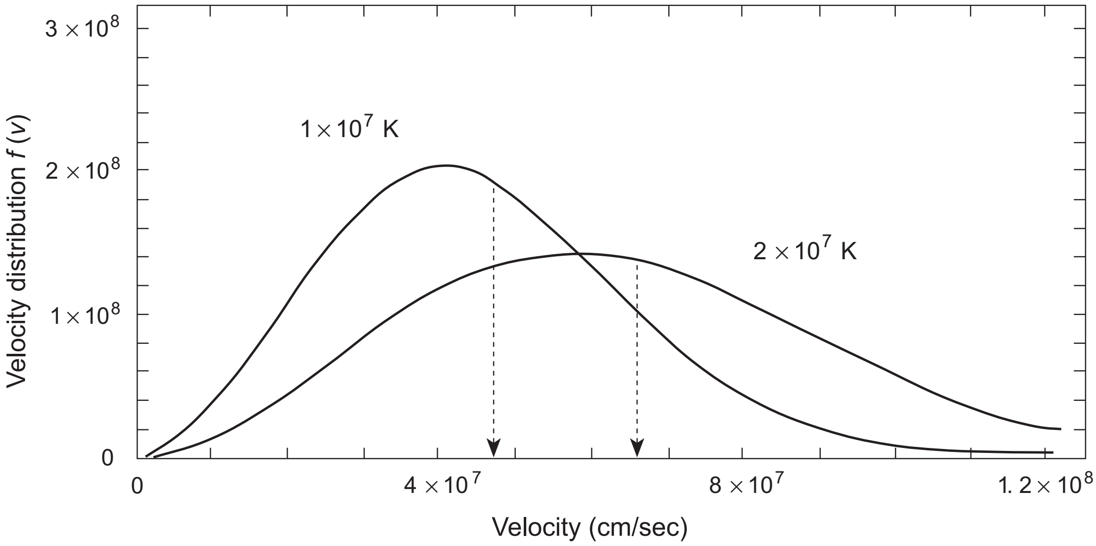


Figure 5.4: Maxwellian velocity distribution for two temperatures; the dashed arrows indicate the mean velocities for each distribution.

5.4 Thermally-Averaged Reaction Rates

The preceding equations assume a monoenergetic beam in a nuclear physics laboratory.

- In a stellar environment we instead have a *gas in approximate thermal equilibrium.*
- If the gas can be described classically, at equilibrium it has a *Maxwell–Boltzmann distribution* $\psi(E)$ of energies

$$\psi(E) = \frac{2}{\pi^{1/2}} \frac{E^{1/2}}{(kT)^{3/2}} \exp(-E/kT).$$

This distribution is illustrated for two different temperatures in Fig. 5.4 as a function of $v = \sqrt{2E/\mu}$.

- A *thermally-averaged cross section* $\langle \sigma v \rangle_{\alpha\beta}$ results from *averaging the cross section over the velocities* in the gas,

$$\begin{aligned}\langle \sigma v \rangle_{\alpha\beta} &\equiv \int_0^\infty \psi(E)\sigma_{\alpha\beta}(E) v\, dE, \\ &= \sqrt{\frac{8}{\pi\mu}}(kT)^{-3/2}\int_0^\infty \sigma_{\alpha\beta}(E)e^{-E/kT}E\, dE,\end{aligned}$$

  where the second form follows from $v = \sqrt{2E/\mu}$.

- Units of $\langle \sigma v \rangle_{\alpha\beta}$ are cm^3 s^{-1} (cross section × velocity).

- We then introduce a *thermally-averaged reaction rate:*

$$\begin{aligned}r_{\alpha\beta} &= n_\alpha n_X \int_0^\infty \psi(E)\sigma_{\alpha\beta}(E) v\, dE = n_\alpha n_X \langle \sigma v \rangle_{\alpha\beta} \\ &= \rho^2 N_{\rm A}^2 \frac{X_\alpha X_X}{A_\alpha A_X}\langle \sigma v \rangle_{\alpha\beta} = \rho^2 N_{\rm A}^2 Y_\alpha Y_X \langle \sigma v \rangle_{\alpha\beta},\end{aligned}$$

  where in the interest of compact notation

  1. We drop explicit display of the δ-function factor necessary when the colliding particles are identical, and
  2. In the second line we introduce the *mass fractions* X_i and the *abundances* Y_i,

$$X_i = \frac{n_i A_i}{\rho N_{\rm A}} \qquad Y_i \equiv \frac{X_i}{A_i} = \frac{n_i}{\rho N_{\rm A}}.$$

- The units of $r_{\alpha\beta}$ are cm^{-3} s^{-1} (*rate per unit volume*),

- The clear *physical interpretation* of

$$r_{\alpha\beta} = n_\alpha n_X \langle \sigma v \rangle_{\alpha\beta}$$

is that the *total rate* for the 2-body reaction

$$\alpha + \mathrm{X} \to \mathrm{Y} + \beta,$$

is given by

  - the (thermally averaged) *rate* $\langle \sigma v \rangle_{\alpha\beta}$ *for a single* α *to react with a single* X to produce $Y + \beta$,
  - multiplied by the *number of* α *per unit volume* n_α, and
  - multiplied by the *number of* X *per unit volume* n_X.

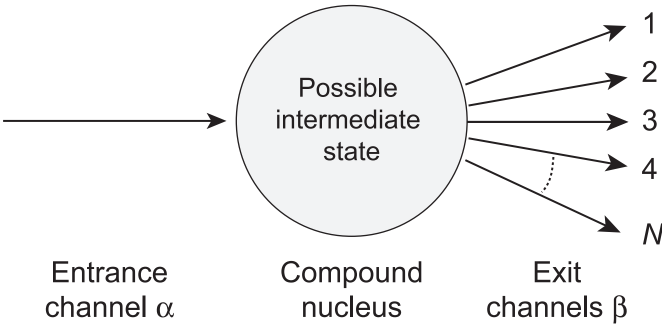


Figure 5.5: Reaction channels. The reaction might or might not involve an intermediate compound nucleus.

5.5 Parameterization of Cross Sections

To proceed we need the *cross section* (in the center of mass system) to calculate the thermally-averaged rates.

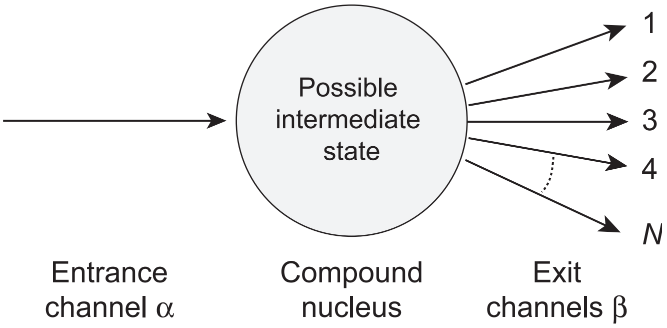
- The cross section may be *parameterized* in the general form

$$\sigma_{\alpha\beta}(E) = \pi g \lambda\!\!\!^{-2} \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma^2} f(E),$$

  where the energy widths $\Gamma_i \equiv \hbar/\tau_i$ are expressed in terms of the corresponding *mean life* τ_i for decay of the compound system through channel i, and

- the *entrance channel* is denoted by α,

- the *exit channel* is denoted by β.



$$\sigma_{\alpha\beta}(E) = \pi g \lambda\!\!\!^{-2} \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma^2} f(E),$$

- the *total width* is $\Gamma = \sum_i \Gamma_i$, where the sum is over all open channels i,
- the *probability to decay to channel* i is $P_i = \Gamma_i/\Gamma$,
- the *reduced deBroglie wavelength* is defined through $\lambda\!\!\!^{-2} = \hbar^2/2\mu E$,
- the *statistical factor* g contains information on the spins of projectile, target, and compound nucleus (*typically of order 1*), and
- the *detailed reaction information* resides in $f(E)$.

The parameters Γ appearing in

$$\sigma_{\alpha\beta}(E) = \pi g \lambda\!\!\!^{-2} \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma^2} f(E)$$

have the units of $\hbar$ *divided by time*, which is *energy*.

- They are called *energy widths* because
  - states with *short lifetimes for decay* (large decay rates) correspond to spectral peaks (resonances) *broad in energy*, by a

    $$\Delta E \cdot \Delta t \simeq \hbar$$

    uncertainty principle argument.
  - Conversely, states with *long decay lifetimes* (small decay rates) correspond to *narrow resonances*.
  - The limiting case is a *completely stable state*, with vanishing decay rate and a *sharply-defined energy*.
- The factor $f(E)$ is generally either
  1. *resonant*, if it is *strongly peaked in energy* because of a narrow (quasibound) compound-nucleus state, or
  2. *nonresonant*, because the reaction energy is far from a resonance, or there are no resonances (quasibound states) in the channel of interest.
- The total rates will typically be a *sum of contributions* for resonant and nonresonant pieces.

5.6 Nonresonant Cross Sections

Most reactions in stellar energy production are *exothermic*,

$$\begin{aligned} Q &\equiv (\text{mass of reactants}) - (\text{mass of products}) \\ &= m_\alpha c^2 + m_\mathrm{X} c^2 - m_\beta c^2 - m_\mathrm{Y} c^2 > 0. \end{aligned}$$

- Typical Q-values for the reactions of interest are ~ 1 MeV.

> This additional energy leads to a *marked asymmetry in the entrance and exit channels* for charged particle reactions.

- In the entrance channel, the *thermal energies available* are set by the temperatures through

  $$kT = 8.6174 \times 10^{-8}\, T\ \text{keV},$$

  with the temperature expressed in kelvin (K).

- *Hydrogen burning* typically occurs in a temperature range

  $$10^7\ \text{K} < T < 10^9\ \text{K},$$

  implying a range of *kinetic energies in the plasma*,

  $$1\ \text{keV} < kT < 100\ \text{keV}.$$

- Thus, for average $Q \sim 1$ MeV, we often have

  $$E(\text{entrance}) \ll E(\text{exit}),$$

  for reactions of interest.

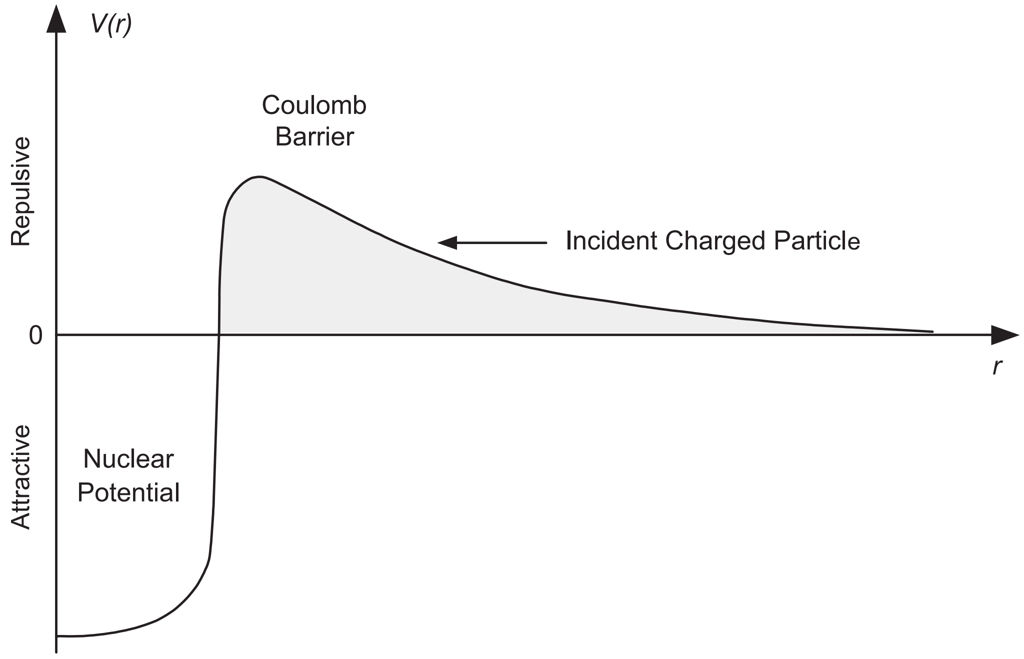


Figure 5.6: The Coulomb barrier for charged-particle reactions.

5.6.1 Coulomb Barriers

Because of *low entrance-channel energies*, charged-particle reactions are *strongly influenced by the Coulomb barrier*

$$E_{\rm CB} = 1.44 \frac{Z_\alpha Z_{\rm X}}{R({\rm fm})}\,{\rm MeV},$$

where Z_i is the atomic number of particle i, the separation R is

$$R \simeq 1.3(A_\alpha^{1/3} + A_{\rm X}^{1/3})\,{\rm fm},$$

- A_i is the atomic mass number (in amu) of particle i,
- energies are in MeV, and
- distances in fermis (fm):$1\,{\rm fm} = 10^{-13}\,{\rm cm} = 10^{-15}\,{\rm m}$).

Example: Consider a proton scattering from a ^{28}Si nucleus. From

$$E_{\text{CB}} = 1.44\,\frac{Z_\alpha Z_{\text{X}}}{R(\text{fm})}\,\text{MeV} \qquad R \simeq 1.3(A_\alpha^{1/3} + A_{\text{X}}^{1/3})\ \text{fm}$$

with

$$Z_\alpha = 1 \qquad Z_X = 14 \qquad A_\alpha = 1 \qquad A_X = 28$$

we obtain

$$E_{\text{CB}} = 1.44\,\frac{(1)(14)}{1.3(1^{1/3} + 28^{1/3})}\,\text{MeV} = 3.8\,\text{MeV}.$$

Table 5.1: Coulomb barriers for p + X

X	$Z_\alpha Z_\beta$	R (fm)	$E_{\rm CB}$ (MeV)
$^1_1$H	1	2.6	0.55
$^{12}_{6}$C	6	4.3	2.0
$^{28}_{14}$Si	14	5.2	3.8
$^{56}_{26}$Fe	26	6.3	6.0

Some *typical Coulomb barriers* for proton reactions p + X are shown in Table 5.1, where we note that

- *Entrance channel energies for hydrogen fusion* in stars (10^{-3} to 10^{-1} MeV) are typically *orders of magnitude lower than the Coulomb barrier.*

  As we shall see, this implies a *dramatic temperature dependence* for hydrogen fusion reactions.

- On the other hand, exit channel energies (approximately 1 MeV in typical cases) are *comparable to the barrier energies* for fusion of protons with lighter ions.

5.6.2 Barrier Penetration Factors

Energies in a stellar plasma are *too small to surmount the Coulomb barrier* for typical charged-particle reactions.

- However, *quantum tunneling* can occur for energies below the height of the barrier, albeit with small probability.
- Assuming $E_{\mathrm{CB}} \gg E$, the barrier penetration probability for a collision with zero relative orbital angular momentum (*s-waves in scattering theory*) is

  $$P(E) \propto e^{-2\pi\eta},$$

  where the dimensionless *Sommerfeld parameter* η is

  $$\eta = \frac{Z_\alpha Z_X e^2}{\hbar v} = \sqrt{\frac{\mu}{2}} \frac{Z_\alpha Z_X e^2}{\hbar E^{1/2}}.$$

- Realistically, $P(E) < \sim \exp(-12)$.
- Thus, charged-particle reactions in stars are *highly-improbable events.*
- The reaction probability will be *dominated by the probability to penetrate the barrier*.
- Thus, we take as an *entrance channel width* for nonresonant reactions

  $$\Gamma_\alpha \simeq e^{-2\pi\eta},$$

  which clearly has a *strong energy dependence.*

5.6.3 Astrophysical S-Factors

Exit channel energies are comparable to barrier energies.

- Thus we assume that Γ_β is a weakly varying function of E, as is $\Gamma = \Gamma_\alpha + \Gamma_\beta$, and
- we express the *nonresonant cross section* as

$$\sigma_{\alpha\beta}(E) = \pi g \lambda\!\!\!^{-2} \frac{\Gamma_\alpha \Gamma_\beta}{\Gamma^2} f(E) \equiv \frac{S(E)}{E} e^{-2\pi\eta},$$

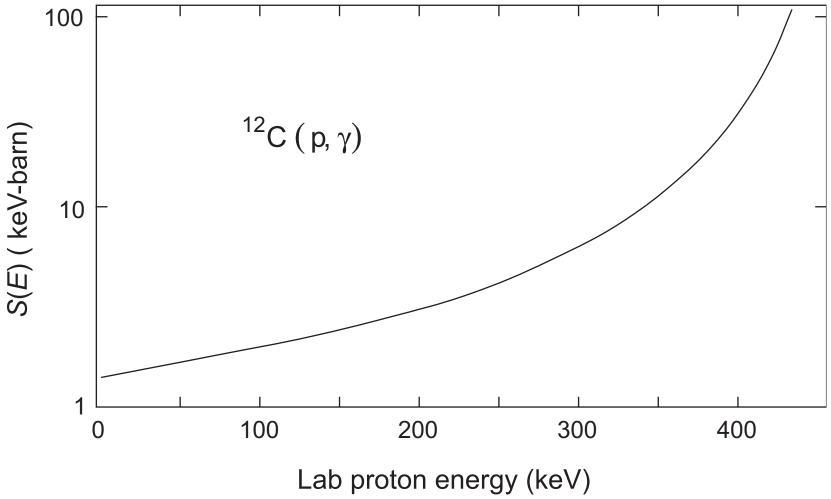
where $\lambda\!\!\!^{-2} \propto 1/E$ and $\Gamma_\alpha \propto e^{-2\pi\eta}$ have been used.

- The *astrophysical S-factor* is defined by

$$S(E) \equiv \sigma_{\alpha\beta}(E) E e^{2\pi\eta}.$$

$S(E)$ varies slowly with E and contains all energy dependence other than $\lambda\!\!\!^{-2}$ or $\exp(-2\pi\eta)$.

The S-factor for a (p, γ) reaction is illustrated below.



Because of the *low energies* $\sim kT$ for stellar plasmas,

- experimental measurements often must be *extrapolated to lower energy*.
- This is usually done by *assuming no resonance* at the lower energy and plotting

  $$S(E) \equiv \sigma_{\alpha\beta}(E)Ee^{2\pi\eta},$$

  which has *smooth behavior* and therefore is *more easily extrapolated* than the full cross section

Then from

$$\langle\sigma v\rangle_{\alpha\beta} = \sqrt{\frac{8}{\pi\mu}}(kT)^{-3/2}\int_0^\infty \sigma_{\alpha\beta}(E)e^{-E/kT}E\,dE,$$

$$\sigma_{\alpha\beta}(E) = \pi g \lambda\!\!\!^{-2}\frac{\Gamma_\alpha\Gamma_\beta}{\Gamma^2}f(E) \equiv \frac{S(E)}{E}e^{-2\pi\eta},$$

the *thermally-averaged nonresonant cross section* is

$$\begin{aligned}\langle\sigma v\rangle_{\alpha\beta} &= \sqrt{\frac{8}{\pi\mu}}(kT)^{-3/2}\int_0^\infty S(E)e^{-E/kT-2\pi\eta}\,dE\\ &= \sqrt{\frac{8}{\pi\mu}}(kT)^{-3/2}\int_0^\infty S(E)e^{-E/kT}e^{-bE^{-1/2}}\,dE,\end{aligned}$$

where we define

$$b \equiv 2\pi\left(\frac{\mu}{2}\right)^{1/2}\frac{Z_\alpha Z_X e^2}{\hbar},$$

and $S(E)$ is in erg cm^2.

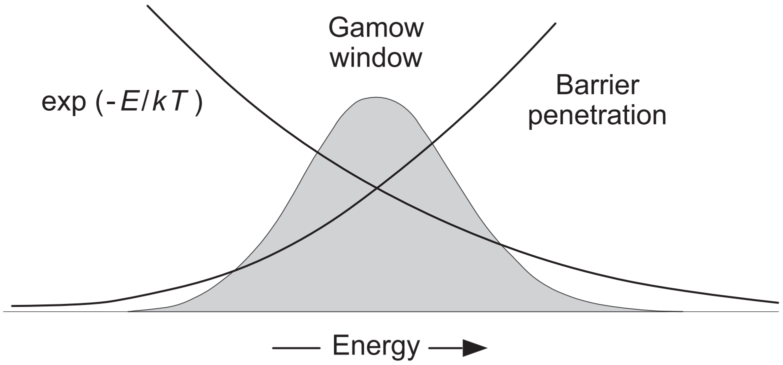


Figure 5.7: The Gamow window.

5.6.4 The Gamow Window

The energy dependence of

$$\langle \sigma v \rangle_{\alpha\beta} = \sqrt{(8/(\pi\mu)}(kT)^{-3/2} \int_0^\infty S(E) \underbrace{e^{-E/kT} e^{-bE^{-1/2}}}_{} dE$$

resides primarily in the factor

$$F_{\mathrm{G}} \equiv e^{-E/kT} e^{-bE^{-1/2}},$$

which is termed the *Gamow window*.

- The first factor $\exp(-E/kT)$, arising from the thermal velocity distribution, *decreases rapidly with energy.*
- The second factor $\exp(-bE^{-1/2})$, arising from the barrier penetration factor, *increases rapidly with energy.*
- Thus, the product is *strongly localized in energy* (Fig. 5.7).

Only for energies within the Gamow window are stellar charged-particle reactions likely to occur.

The *maximum of the Gamow peak* (*Problem 5.4 ****)

$$E_0 = 1.22(Z_\alpha^2 Z_\mathrm{X}^2 \mu T_6^2)^{1/3}\,\mathrm{keV}.$$

- For many reactions of interest this is *only tens of keV*.
- Hence laboratory cross sections must be *extrapolated to these low energies* to calculate astrophysical processes,
- because it is very *difficult to do reliable experiments at such low energies*.

Useful approximate expressions can be obtained by *assuming the Gamow peak to be a Gaussian* (*Problem 5.5 ****).

- In this approximation the *width of the Gamow peak* is

  $$\Delta = \frac{4}{3^{1/2}}(E_0 kT)^{1/2} = 0.75(Z_\alpha^2 Z_X^2 \mu T_6^5)^{1/6}\,\mathrm{keV}$$

  and the *cross section* is

  $$\langle \sigma v \rangle_{\alpha\beta} \simeq \frac{0.72\times 10^{-18} S(E_0) a^2}{\mu Z_\alpha Z_\mathrm{X} T_6^{2/3}} \exp(-a T_6^{-1/3}),$$

  in units of cm^3 s^{-1}, where

  $$a = 42.49(Z_\alpha^2 Z_\mathrm{X}^2 \mu)^{1/3},$$

  and $S(E_0)$ is evaluated at the energy of the Gamow peak in units of keV barns.

Example: From the gaussian approximation we find that for the interaction of two protons at a temperature of $T_6 = 20$ (that is, $T = 20 \times 10^6\,\mathrm{K}$),

$$kT = 1.7\,\mathrm{keV} \qquad E_0 = 7.2\,\mathrm{keV} \qquad \Delta = 8.2\,\mathrm{keV},$$

and from the earlier table the corresponding Coulomb barrier is about 500 keV.

5.7 Resonant Cross Sections

In the simplest case of an *isolated resonance*,

- $f(E)$ can be expressed in the *Breit–Wigner form*

$$f(E)_{\mathrm{res}} = \frac{\Gamma^2}{(E-E_{\mathrm{r}})^2 + (\Gamma/2)^2},$$

- where the *resonance energy* E_{r} is related to a corresponding *excitation energy* E^* *for a quasibound state* in the compound nucleus through

$$E_{\mathrm{r}} = (m_{\mathrm{Z}} - m_\alpha - m_{\mathrm{X}})c^2 + E^*.$$

The corresponding *Breit–Wigner cross section* is

$$\sigma_{\alpha\beta} = \pi g \lambda\!\!\!^{-2} \frac{\Gamma_\alpha \Gamma_\beta}{(E-E_{\mathrm{r}})^2 + (\Gamma/2)^2},$$

and will generally exhibit a *strong peak near* $E = E_{\mathrm{r}}$.

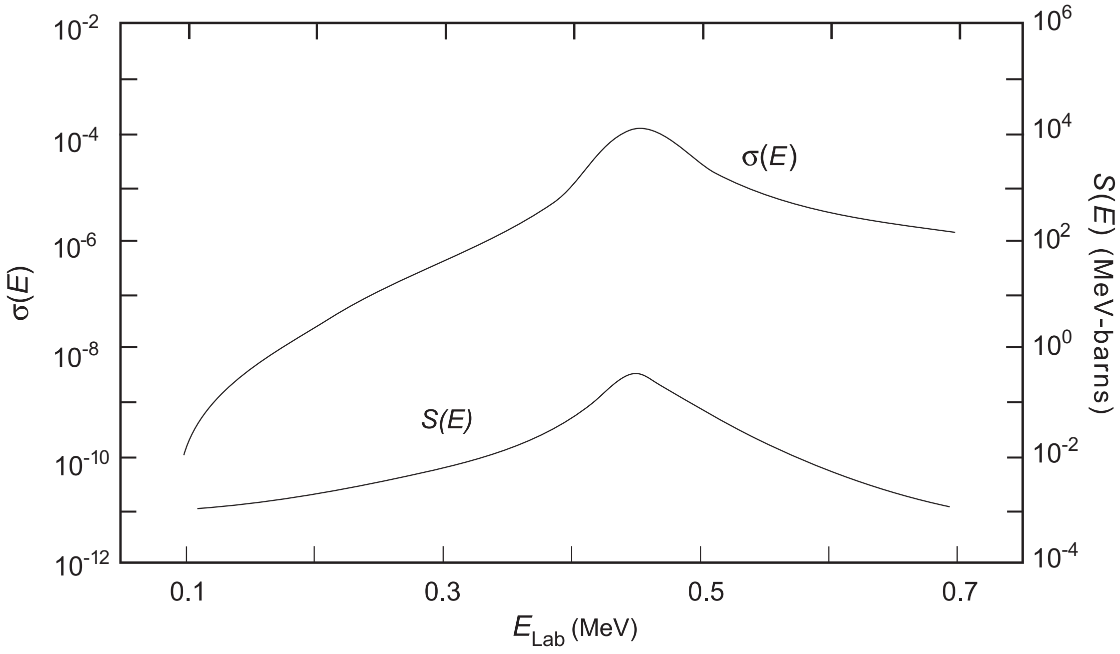


Figure 5.8: Cross section $\sigma(E)$ in barns for the reaction $^{12}\mathrm{C}(\mathrm{p},\gamma)^{13}\mathrm{N}$.

Example: Consider the reaction $^{12}\mathrm{C}(\mathrm{p},\gamma)^{13}\mathrm{N}$ illustrated in Fig. 5.8.

- This reaction has a resonance corresponding to a state in $^{13}\mathrm{N}$ at an excitation energy of 2.37 MeV.

- Thus it is strongly excited at a laboratory proton energy of 0.46 MeV.

If the Maxwell–Boltzmann distribution $\psi(E)$ and the widths Γ_i vary slowly over a resonance, we may assume

$$\psi(E) \to \psi(E_\mathrm{r}) \qquad \Gamma_\alpha \to \Gamma_\alpha(E_\mathrm{r}) \qquad \Gamma_\beta \to \Gamma_\beta(E_\mathrm{r}),$$

and we obtain he resonant velocity-averaged cross section

$$\begin{aligned}\langle\sigma v\rangle_{\alpha\beta} &= \frac{\pi\hbar^2 g}{2\mu}\sqrt{\frac{8}{\pi\mu}}(kT)^{-3/2}e^{-E_\mathrm{r}/kT} \\ &\quad\times \Gamma_\alpha(E_\mathrm{r})\Gamma_\beta(E_\mathrm{r})\int_0^\infty \frac{1}{(E-E_\mathrm{r})^2+(\Gamma/2)^2}\,dE.\end{aligned}$$

If the resonance is broad or E_r is small, the preceding assumptions may be invalid and it may be necessary to integrate over the resonance energy numerically using the data.

The integrand peaks near E_r; extending the lower limit of the integral to $-\infty$ and assuming the widths to be constant gives

$$\langle\sigma v\rangle_{\alpha\beta} = 2.56\times 10^{-13}\frac{(\omega\gamma)_\mathrm{r}}{(\mu T_9)^{3/2}}\exp(-11.605E_\mathrm{r}/T_9)\,\mathrm{cm}^3\,\mathrm{s}^{-1},$$

where E_r is in MeV, T_9 indicates temperature in units of 10^9 K, and

$$(\omega\gamma)_\mathrm{r} \equiv g\frac{\Gamma_\alpha\Gamma_\beta}{\Gamma}$$

has units of MeV and is tabulated for reactions of interest.

5.8 Libraries of Cross Sections

The Gamow window is *not Gaussian*.

- By using expansions to characterize the deviation from Gaussian behavior of the realistic curve,
- *correction terms* may be derived that give a more accurate representation of the thermally-averaged cross section.
- One parameterization that incorporates such correction terms and is often used in reaction rate compilations is

$$\langle \sigma v \rangle = a(f_0 + f_1 T^{1/3} + f_2 T^{2/3} + f_3 T + f_4 T^{4/3} + f_5 T^{5/3}) \frac{e^{-bT^{-1/3}}}{T^{2/3}}$$

  where a, b, and f_n parameterize the cross section.
- The *Caughlan and Fowler compilation* used in some problems in the book is parameterized in this manner.
- Another reaction library used for rates in many of our examples is *ReacLib*, which is described in *Appendix D.2 of the book*.

5.9 Total Rate of Energy Production

- The *total reaction rate per unit volume* $r_{\alpha\beta}$ is given by

$$r_{\alpha\beta} = \rho^2 N_{\rm A}^2 \frac{X_\alpha X_{\rm X}}{A_\alpha A_{\rm X}} \langle \sigma v \rangle_{\alpha\beta} = \rho^2 N_{\rm A}^2 Y_\alpha Y_{\rm X} \langle \sigma v \rangle_{\alpha\beta}.$$

- The corresponding *total rate of energy production per unit mass* is then given by the product of the rate and the Q-value, divided by the density:

$$\varepsilon_{\alpha\beta} = \frac{r_{\alpha\beta} Q}{\rho},$$

  which has CGS units of erg g^{-1}s^{-1}.

- The Q-value entering this expression is defined by

$$Q \equiv (\text{mass of reactants}) - (\text{mass of products})$$

  but with the proviso that

> If a reaction produces a *neutrino that removes energy from the star* without appreciable interaction, its energy should be subtracted from the total Q-value.

5.10 Temperature and Density Exponents

It is often useful to parameterize the energy production rate of a star in the *power-law form*

$$\varepsilon = \varepsilon_0 \rho^{\lambda} T^{\nu}.$$

- Then the behavior of the energy production may be characterized in terms of
  - the *temperature exponent* ν and
  - the *density exponent* λ.
- The energy production is *not universally of this form*.
- However, this approximation with constant exponents is usually *valid for a limited range of* T *and* ρ.
- Energy production mechanisms for stars are often *operative only in narrow ranges of temperature and density*.
- Hence the exponents λ and ν can provide a useful parameterization for the regions of physical interest.

We may define temperature and density exponents for an *arbitrary energy production function* $\varepsilon(\rho, T)$ through

$$\lambda = \left(\frac{\partial \ln \varepsilon}{\partial \ln \rho}\right)_T \qquad \nu = \left(\frac{\partial \ln \varepsilon}{\partial \ln T}\right)_\rho .$$

Table 5.2: Density and temperature exponents

Stellar process	Density (λ)	Temperature (ν)
PP chain	1	~ 4
CNO cycle	1	~ 16
Triple-α	2	~ 40

Temperature and density exponents are displayed in Table 5.2 for the PP chain and CNO cycle, and for the triple-α process that burns helium to carbon in red giant stars.

- Notice in Table 5.2 the exquisite temperature dependence exhibited by these reactions.
- This *enormous sensitivity of energy production to temperature* is a central feature of stellar structure and stellar evolution.

5.11 Reaction Selection Rules

Sometimes it is possible to infer the astrophysical significance of various nuclear reactions based on *selection rules and conservation laws,* without having to calculate any detailed rates.

- *Angular momentum* is conserved in all reactions.
- Thus the angular momentum $\boldsymbol{J}$ of a compound nucleus state populated in a two-body reaction must satisfy

$$\boldsymbol{j}_1 + \boldsymbol{j}_2 + \boldsymbol{l} = \boldsymbol{J}.$$

  where

  - $\boldsymbol{j}_1$ and $\boldsymbol{j}_2$ are the angular momenta associated with the colliding particles and
  - $\boldsymbol{l}$ is the angular momentum of relative orbital motion in the entrance channel.

- Likewise, *isotopic spin* (an abstract approximate symmetry) is conserved to a high degree in strong interactions.
- Thus the isotopic spins in a two-body reaction must approximately satisfy

$$\boldsymbol{t}_1 + \boldsymbol{t}_2 = \boldsymbol{T},$$

  where

  - $\boldsymbol{t}_1$ and $\boldsymbol{t}_2$ are the isotopic spins associated with the colliding particles and
  - $\boldsymbol{T}$ is the isotopic spin of the final state populated.

- *Parity* (symmetry of the wavefunction under *space reflection*)
  - is *maximally broken* in the weak interactions, but
  - is *conserved* in the strong and electromagnetic reactions.

- In a nuclear reaction that does not involve the weak force, the *parities must satisfy*

  $$(-1)^l \pi(j_1)\pi(j_2) = \pi(J).$$

  where

  - $\pi = \pm$ denotes the *parity* and
  - j_i the *angular momentum* of the states.

Compound nucleus states with angular momentum, isospin, and parity quantum numbers that do not satisfy these selection rules will generally not be populated significantly in reactions.

For nuclei with even numbers of protons and neutrons *(even–even nuclei)*

- The ground states always have angular momentum and parity $J^\pi = 0^+$.
- Therefore, if the colliding nuclei are
  - even–even nuclei and
  - in their ground states,
- the angular momentum and the parity of the state excited in the compound nucleus are *both* determined completely by the orbital angular momentum l of the entrance channel:
$$J = l \qquad \pi(J) = (-1)^l.$$
- Resonance states satisfying this condition are said to have *natural parity.*

Example: Consider the reaction $\alpha + {}^{16}\text{O} \rightarrow {}^{20}\text{Ne}^*$ (where the * on the Ne indicates that it is in an excited state).

- Under normal conditions the α-particle and ^{16}O will be in their ground states and thus will each have $J^\pi = 0^+$.
- Therefore, parity conservation requires that any state excited in ^{20}Ne by this reaction have parity

$$\pi({}^{20}\text{Ne}) = (-1)^l = (-1)^J.$$

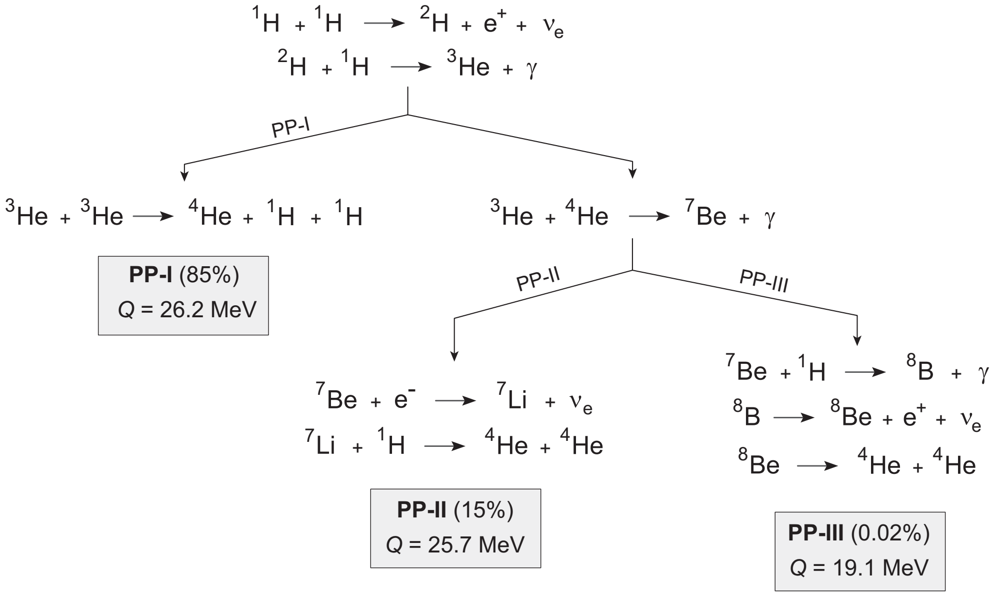
- We conclude that
  - states in ^{20}Ne having $J^\pi = 0^+, 1^-, 2^+, 3^-, \ldots$ *may be populated* (because they are natural parity) but that
  - population of states having (say) $J^\pi = 2^-$ or 3^+ is *forbidden*, or at least strongly suppressed (not natural parity).
- In the ^{20}Ne spectrum there is a state at 4.97 MeV of excitation relative to the ground state having $J^\pi = 2^-$.
- This state *cannot be excited* in the capture reaction $\alpha + {}^{16}\text{O} \rightarrow {}^{20}\text{Ne} + \gamma$ because it is *not a natural parity state*.

As we shall see in the later discussion of helium burning, *if this seemingly obscure state had the opposite parity, we would not exist (!).*

Chapter 6

Stellar Burning Processes

The *slowest reaction* in the PP chain,



$^1H + {}^1H \longrightarrow {}^2H + e^+ + \nu_e$
$^2H + {}^1H \longrightarrow {}^3He + \gamma$
PP-I
$^3He + {}^3He \longrightarrow {}^4He + {}^1H + {}^1H$
PP-I (85%)
Q = 26.2 MeV
$^3He + {}^4He \longrightarrow {}^7Be + \gamma$
PP-II
PP-III
$^7Be + e^- \longrightarrow {}^7Li + \nu_e$
$^7Li + {}^1H \longrightarrow {}^4He + {}^4He$
PP-II (15%)
Q = 25.7 MeV
$^7Be + {}^1H \longrightarrow {}^8B + \gamma$
$^8B \longrightarrow {}^8Be + e^+ + \nu_e$
$^8Be \longrightarrow {}^4He + {}^4He$
PP-III (0.02%)
Q = 19.1 MeV

and therefore the one that *governs the overall rate* at which the chain produces power, is the initial step

$$^1H + {}^1H \longrightarrow {}^2H + e^+ + \nu_e.$$

6.1 Reactions of the Proton–Proton Chain

The reaction ${}^{1}\mathrm{H} + {}^{1}\mathrm{H} \longrightarrow {}^{2}\mathrm{H} + \mathrm{e}^{+} + \nu_{\mathrm{e}}$

- is *very slow* because it proceeds by *weak interactions* (the *neutrino* indicates this).
- It is also *non-resonant*.
- The reaction rate is found to be

$$\begin{aligned} r_{\mathrm{pp}} &= \tfrac{1}{2} n_{\mathrm{p}}^{2} \langle \sigma v \rangle_{\mathrm{pp}} \\ &= \frac{1.15 \times 10^{9}}{T_9^{2/3}} X^2 \rho^2 \exp(-3.38/T_9^{1/3})\ \mathrm{cm}^{-3}\ \mathrm{s}^{-1}, \end{aligned}$$

  where X is the hydrogen mass fraction.
- The temperature exponent is (Problem)

$$\nu_{\mathrm{pp}} = 11.3/T_6^{1/3} - 2/3,$$

  implying that $\nu_{\mathrm{pp}} \simeq 4$ for $T_6 = 15$.

- The *rate of change for proton number* because of this reaction is given by the usual radioactive decay law,

  $$\frac{dn}{dt} = -\frac{1}{\tau}n,$$

  where τ is the mean life for the reaction.

- Thus, the *mean life for a proton* with respect to being converted in the PP chain is

  $$\tau_{\mathrm{p}} = -\frac{n_{\mathrm{p}}}{dn_{\mathrm{p}}/dt} = \frac{n_{\mathrm{p}}}{2r_{\mathrm{pp}}},$$

  where a factor of two appears in the denominator because *two protons are destroyed in each reaction*.

Example: For *typical solar conditions* we may take a temperature of $T_6 = 15$, a central density of $\rho = 100\,\mathrm{g\,cm^{-3}}$, and a central hydrogen mass fraction of $X = 0.5$.

- Then the preceding equations yield an estimate of

  $$\tau_{\mathrm{p}} \simeq 6\times 10^9\ \mathrm{years}.$$

- This is remarkably long and sets the scale for the main sequence life of the Sun.

The slowness of the initial step in the PP-chain is ultimately because *the diproton (^{2}He) is not a bound system*.

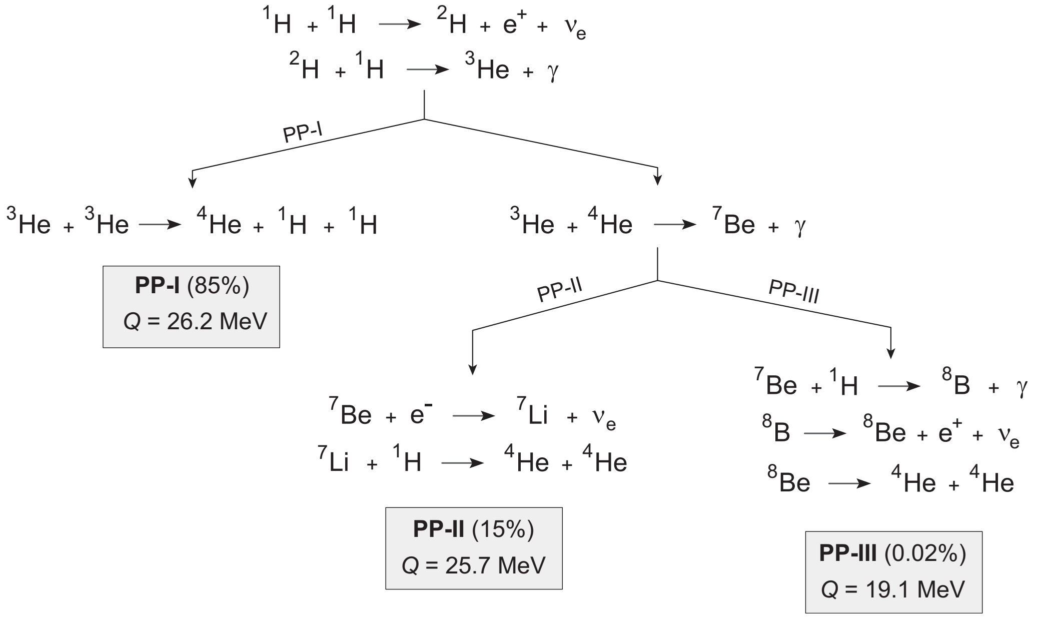
- If the diproton were bound,
  - the first step of the PP-chain could be a strong interaction and
  - the lifetime would be much shorter.
- Instead, the first step must wait for a *highly improbable event:* a weak decay of a proton from a broad *p–p* resonance having a very short lifetime.
- In contrast, the mean life for the deuterium produced in the first step and consumed in the next step

  $$\mathrm{p} + \mathrm{d} \rightarrow {}^3\mathrm{He} + \gamma$$

  is about *one minute* under solar conditions.
- The final fusion of two helium-3 isotopes to form helium-4 is much slower ($\tau \sim 10^6$ years), but is *orders of magnitude faster than the first step*.

Thus, the initial step of the PP chain

- *governs the rate* of the reaction and in turn
- *sets the main sequence lifetime* for stars running on the PP chain.


$^1H + ^1H \rightarrow ^2H + e^+ + \nu_e$
$^2H + ^1H \rightarrow ^3He + \gamma$
PP-I
$^3He + ^3He \rightarrow ^4He + ^1H + ^1H$
PP-I (85%)
Q = 26.2 MeV
$^3He + ^4He \rightarrow ^7Be + \gamma$
PP-II
PP-III
$^7Be + e^- \rightarrow ^7Li + \nu_e$
$^7Li + ^1H \rightarrow ^4He + ^4He$
PP-II (15%)
Q = 25.7 MeV
$^7Be + ^1H \rightarrow ^8B + \gamma$
$^8B \rightarrow ^8Be + e^+ + \nu_e$
$^8Be \rightarrow ^4He + ^4He$
PP-III (0.02%)
Q = 19.1 MeV

The relative importance of PP-I versus PP-II and PP-III depends on the *competition between the reactions*

$$^3\mathrm{He}(^3\mathrm{He}, 2\mathrm{p})^4\mathrm{He} \qquad \text{and} \qquad ^3\mathrm{He}(^4\mathrm{He}, \gamma)^7\mathrm{Be}.$$

- For temperatures where PP is important, the first reaction is faster than the second by about four orders of magnitude, ensuring *dominance of PP-I over PP-II and PP-III*.

- The branching between PP-II and PP-III depends on *competition between electron and proton capture on ^{7}Be*.

- At the temperature of the Sun, electron capture dominates and *PP-II is much stronger than PP-III*.

- At somewhat higher temperatures, PP-III will make much larger contributions. (However, at higher temperatures the CNO process will become more important than PP.)

Table 6.1: Effective Q-values

Process	Q_{eff} (MeV)	% Solar energy
PP-I	26.2	83.7
PP-II	25.7	14.7
PP-III	19.1	0.02
CNO	23.8	1.6

Effective Q-values for the PP chain

- depend on *which subchain is followed*, since
- the *energy carried off by neutrinos* is different in the 3 cases.

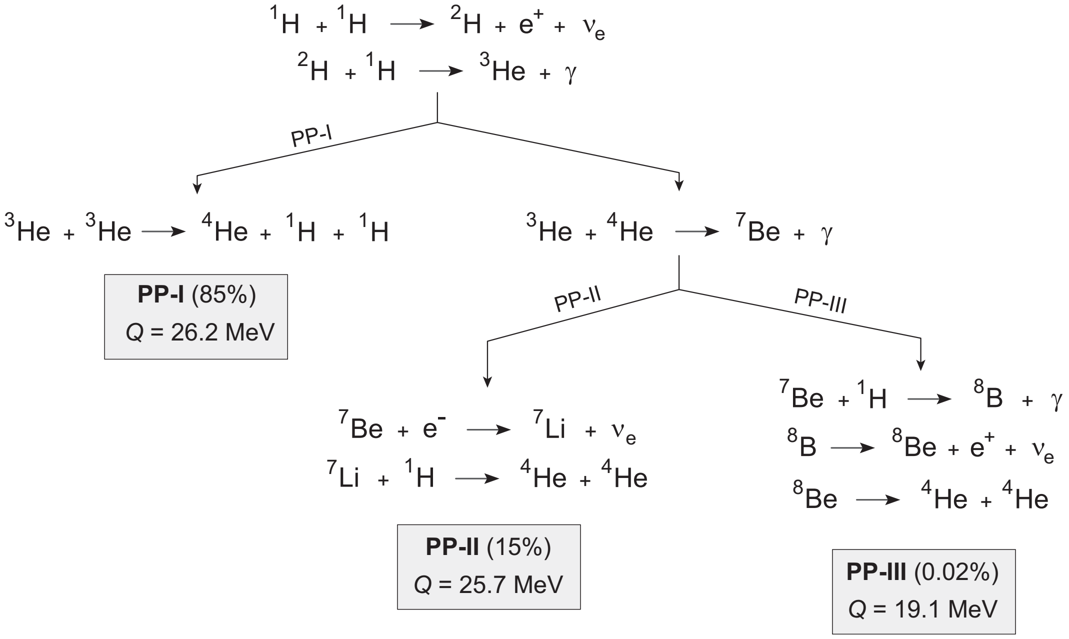
The effective Q-values are listed in Table 6.1.

- Average *energy released per PP chain fusion* in the Sun is

$$\overline{\Delta E}_{\text{pp}} = 0.85\left(\frac{26.2}{2}\right) + (0.15)(25.7) = 15\,\text{MeV},$$

  where

  - *PP-III has been ignored* and
  - the *factor of 2 in the denominator* of the first term results from the first two steps of PP-I needing to run twice to provide two ^{3}He nuclei.


$^1H + {}^1H \rightarrow {}^2H + e^+ + \nu_e$
$^2H + {}^1H \rightarrow {}^3He + \gamma$
PP-I
$^3He + {}^3He \rightarrow {}^4He + {}^1H + {}^1H$
PP-I (85%)
Q = 26.2 MeV
$^3He + {}^4He \rightarrow {}^7Be + \gamma$
PP-II
PP-III
$^7Be + e^- \rightarrow {}^7Li + \nu_e$
$^7Li + {}^1H \rightarrow {}^4He + {}^4He$
PP-II (15%)
Q = 25.7 MeV
$^7Be + {}^1H \rightarrow {}^8B + \gamma$
$^8B \rightarrow {}^8Be + e^+ + \nu_e$
$^8Be \rightarrow {}^4He + {}^4He$
PP-III (0.02%)
Q = 19.1 MeV

- Although PP-III
  - has *negligible influence on energy production*,
  - it *produces much higher energy neutrinos* than PP-I or PP-II.

- The PP-III chain is *highly temperature dependent* because it is initiated by proton capture on a $Z = 4$ nucleus (*Coulomb barrier*).

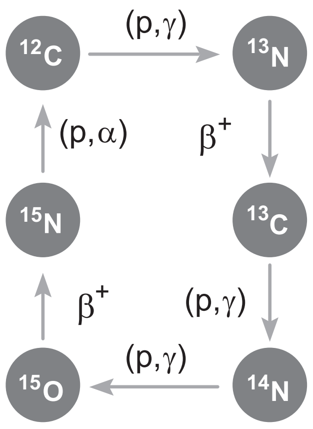
- Therefore, detection of high-energy neutrinos from PP-III
  - can provide a *sensitive probe of the central temperature* of the Sun.
  - We shall return to this issue when we discuss the *solar neutrino problem*.

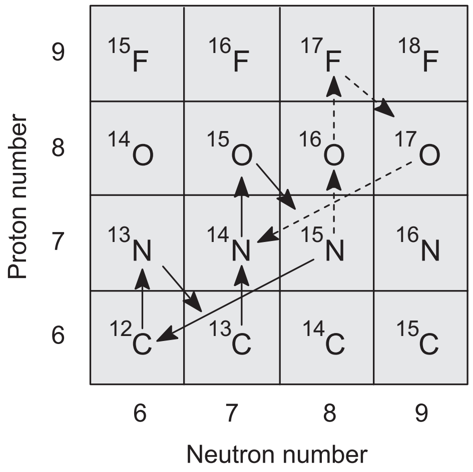
6.2 Reactions of the CNO Cycle

Because of the influence of

- Coulomb barriers and
- *S*-factors,

the *slowest reaction in the CNO cycle,*





is typically found to be

$$p + {}^{14}\text{N} \rightarrow {}^{15}\text{O} + \gamma,$$

which has $S \sim 3.3\,\text{keV barns}$ in the energy range of interest.

- The corresponding *mean life for ^{14}N* against this reaction in the core of the Sun is approximately

$$\tau_{\text{14-N}} \sim 5 \times 10^8 \text{ years.}$$

- The *number density of ^{14}N* at the core of the Sun is

$$n_{\text{14-N}} \sim 2.6 \times 10^{22}\,\text{cm}^{-3},$$

corresponding to an abundance $Y \sim 0.006$.

- The *hydrogen number density* is $n_{\text{H}} \sim 3 \times 10^{25}\,\text{cm}^{-3}$ and
- earlier the *mean life for consumption of a proton* by the PP chain was estimated to be $\tau_{\text{pp}} \sim 6 \times 10^9$ years.
- These numbers imply that the *ratio of PP chain to CNO cycle reactions* in the core of the Sun is approximately

$$\left(\frac{\text{rate for PP}}{\text{rate for } ^{14}\text{N} + \text{p}}\right) \simeq \left(\frac{\tau_{\text{14-N}}}{\tau_{\text{pp}}}\right)\left(\frac{3 \times 10^{25}}{2.6 \times 10^{22}}\right) \simeq 100,$$

- We conclude that for conditions prevailing in the Sun *the PP chain dominates the CNO cycle*.

Detailed calculations indicate that

- the Sun is producing *98.4% of its energy from the PP chain* and
- only *1.6% from the CNO cycle*.

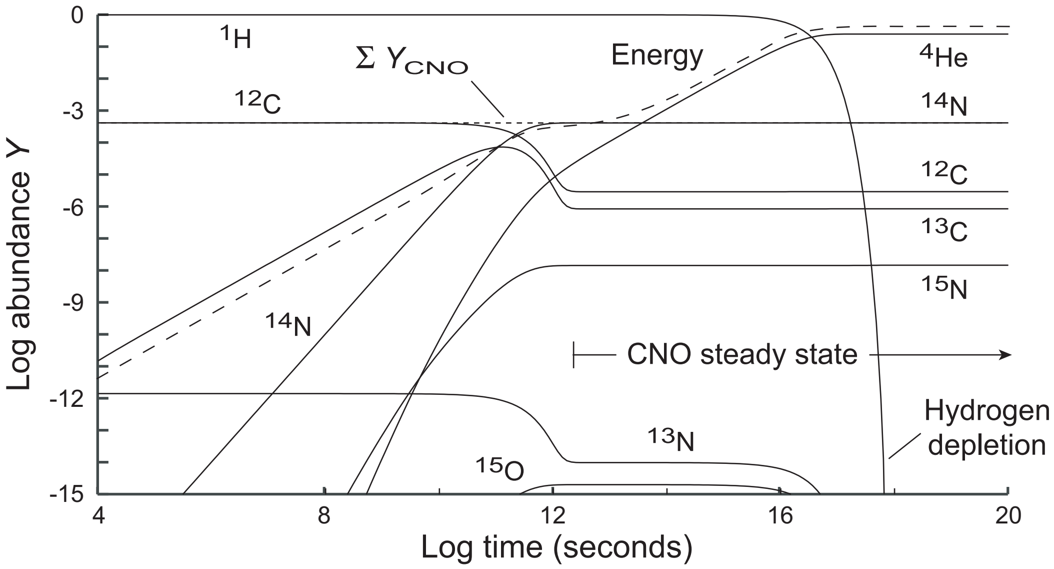
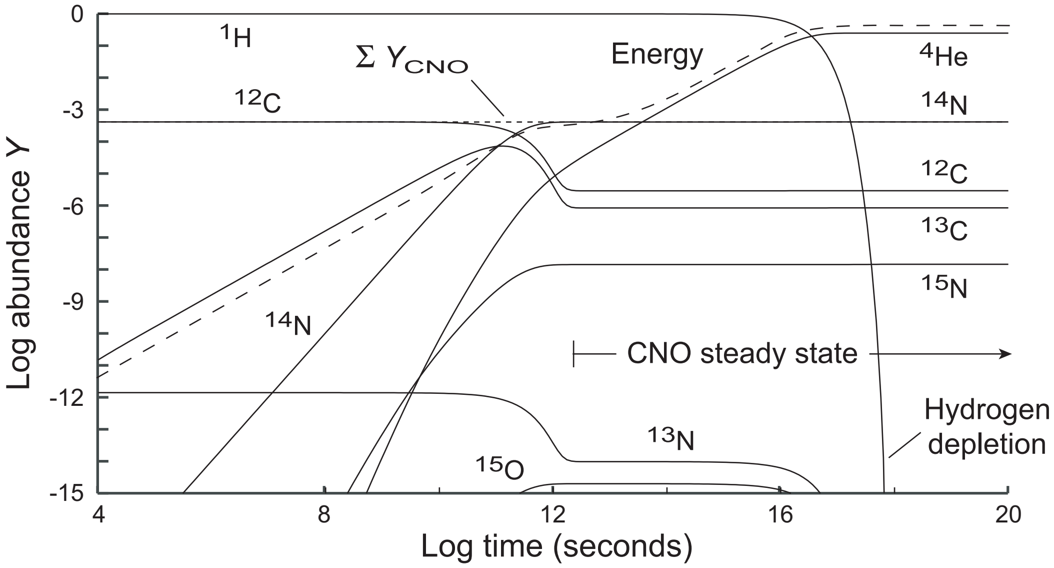


Figure 6.1: CNO cycle run to completion with only hydrogen and a small amount of ^{12}C initially. Mass fractions are shown as solid lines (the mass fractions for ^{13}N and ^{15}O are of order 10^{-14} or smaller and are offscale on this plot). The dashed line is the integrated energy release on an arbitrary log scale. The temperature and density were assumed constant at $T_6 = 20$ and $\rho = 100\,\mathrm{g\,cm^{-3}}$, respectively.

Fig. 6.1 implements a *numerical calculation of CNO abundances* carried to hydrogen depletion for a star with a constant temperature of $T_6 = 20$ and constant density of $100\,\mathrm{g\,cm^{-3}}$.

- An initial mixture of *only two isotopes:* ^{1}H ($X_\mathrm{H} = 0.995$) and ^{12}C ($X_\text{12-C} = 0.005$) was assumed.
- Even though we start with *only a trace amount of one CNO isotope* (^{12}C),
  - the cycle eventually generates an *equilibrium abundance of all CNO isotopes* and
  - steadily releases energy by converting all the H to He.



- Once the cycle is in equilibrium,
  - the *mass fractions of the CNO isotopes remain constant*, so
  - they may be viewed as *catalyzing the conversion of hydrogen to helium.*
- Notice also the result (a quite general one) that
  - the *CNO cycle run to equilibrium produces ^{14}N* as the dominant CNO isotope,
  - even though there was *no initial abundance of ^{14}N* in this particular simulation.

It is thought that *most of the ^{14}N in the Universe* has been produced by the CNO cycle.

- The *effective Q-value* for the CNO cycle is 23.8 MeV.
- The *rate of energy production* is

$$\varepsilon_{\rm CNO} = \frac{4.4 \times 10^{25} \rho XZ}{T_9^{2/3}} \exp(-15.228/T_9^{1/3}) \text{ erg g}^{-1}\text{ s}^{-1},$$

- and the corresponding *temperature exponent* is

$$\nu_{\rm CNO} = 50.8/T_6^{1/3} - 2/3,$$

  which gives

$$\nu_{\rm CNO} \simeq 18 \quad \text{for} \quad T_6 = 20.$$

- This *remarkably strong temperature dependence* implies that

> If the Sun were only slightly hotter, the CNO cycle instead of the PP chain would be the dominant energy production mechanism.

6.3 The Triple-Alpha Process

Main sequence stars produce power by hydrogen fusion. This builds up a thermonuclear ash of helium in the core of the star.

- The star continues to fuse hydrogen to helium in a shell surrounding the central core of helium that is built up.
- This *hydrogen shell burning* adds gradually to the accumulating core of helium and
- The star remains on the main sequence until about 10% of its initial hydrogen has been consumed.
- Fusion of the helium to heavier elements is difficult because
  - There is a *large Coulomb barrier,*
  - There are *no stable mass-5 and mass-8 isotopes* to serve as intermediaries in producing heavier species.
- Thus, helium fusion requires *very high temperatures and densities:*
  - temperatures in excess of about 10^8 K and
  - densities of $10^2 - 10^5\,\mathrm{g\,cm^{-3}}$.

Such conditions can result when stars exhaust their hydrogen fuel and their cores begin to contract.

- Because there are *no stable mass-8 isotopes*, the resulting fusion of helium must involve a *two-step process* in which
  - two helium ions (alpha-particles) combine to form highly unstable ^{8}Be, and
  - this in turn combines with another helium ion to form carbon.
- The resulting sequence,
  - which is *crucial to the power generated by red giant stars* and
  - to the *production of most of the carbon and oxygen* in the Universe,

  is called the *triple-α process.*

Our bodies are composed of about

- 65% oxygen and
- 18% carbon.

We *owe our very existence to the triple-α process*, as we discuss further below!

The burning of helium to carbon by the triple-α process may be viewed as taking place in *three basic steps:*

1. A *small transient population* of ^{8}Be is built up by He + He fusion,
$$^4\mathrm{He} + {}^4\mathrm{He} \leftrightarrow {}^8\mathrm{Be}.$$

2. A *small transient population* of ^{12}C* in an excited state is built up by the reaction
$$^4\mathrm{He} + {}^8\mathrm{Be} \leftrightarrow {}^{12}\mathrm{C}^*.$$

   To produce enough ^{12}C* this reaction *must be resonant,* to compete with $^8\mathrm{Be} \rightarrow {}^4\mathrm{He} + {}^4\mathrm{He}$.

3. A small fraction of the ^{12}C* excited states *decay electromagnetically* by
$$^{12}\mathrm{C}^* \rightarrow {}^{12}\mathrm{C} + 2\gamma$$
to the ground state of carbon-12.

This (highly improbable) sequence of reactions converts three helium ions to ^{12}C,

$$3\alpha \rightarrow {}^{12}\mathrm{C},$$

with an energy release $Q = +7.275\,\mathrm{MeV}$. Let's consider each of these steps in more detail.

6.3.1 Equilibrium Population of Beryllium-8

- The mean life for decay of ^{8}Be back into 2α is $\tau \simeq 10^{-16}$ seconds, which corresponds to a width of

$$\Gamma_{\text{8-Be}} = \hbar\tau^{-1} = 6.8\,\text{eV}.$$

- The capture will be
  - *too slow to compete* with this decay back into 2α
  - unless the corresponding *resonance peak overlaps the Gamow peak*.

- Thus, we expect this rate-controlling step
  - to be significant only if the Gamow energy is *comparable to the Q-value of 92 keV*.
  - This in turn sets the *required conditions for triple-α* to proceed.

- The *maximum of the Gamow peak* is given by

$$E_0 = 1.22(Z_\alpha^2 Z_\text{X}^2 \mu T_6^2)^{1/3}\,\text{keV},$$

implying a temperature of 1.2×10^8 K for $E_0 = 92\,\text{keV}$.

Only for temperatures of order 10^8 K can the initial step of the triple-α reaction produce a sufficient equilibrium concentration of ^{8}Be.

- The preceding simple estimate
  - ignores details such as effects of electron screening,
  - but it sets the correct order of magnitude.
- Also note that this temperature estimate raises the question of why helium was not consumed in big bang nucleosynthesis by the triple-α mechanism. The answer:
  - The temperature was high enough but not the density.
  - Both high temperatures and high densities, which awaited the formation of stars, were required to produce significant amounts of carbon by the triple-α mechanism.

We may estimate the equilibrium concentration of ^{8}Be by application to nuclei of a suitable modification of the atomic Saha equations. The required changes are

1. Replace the number densities of ions and electrons with the number densities of α-particles.
2. Replace the number density of neutral atoms with the number density of ^{8}Be.
3. Replace the statistical factors g for atoms with corresponding statistical factors associated with nuclei.

   This is trivial for the present case: the ground states of both ^{8}Be and ^{4}He have angular momentum zero and $g = 1$ in both cases.

4. The ionization potentials entering the atomic Saha equations are replaced by Q-values in the nuclear case.

   In the present example, $Q = 91.78\,\text{keV}$ for $^8\text{Be} \rightarrow \alpha\alpha$ ("ionization" of ^{8}Be to two α-particles).

5. The electron mass in the atomic Saha equations is replaced by the reduced mass $m_\alpha m_\alpha / 2m_\alpha = \frac{1}{2}m_\alpha$.

The resulting *nuclear Saha equation* is

$$\frac{n_\alpha^2}{n(^8\mathrm{Be})} = \left(\frac{\pi k T m_\alpha}{h^2}\right)^{3/2} \exp(-Q/kT).$$

The situation where such equations are applicable is termed *nuclear statistical equilibrium (NSE)*.

Example: *Helium flashes* are explosive helium burning events that can occur in red giant stars.

- *Typical helium flash conditions* correspond to
  - a temperature of $T_9 \simeq 0.1$ and
  - a density of $\rho \simeq 10^6 \,\mathrm{g\,cm^{-3}}$.
- For a *triple-α powered helium flash* in a pure helium core, we obtain from

$$\frac{n_\alpha^2}{n(^8\mathrm{Be})} = \left(\frac{\pi k T m_\alpha}{h^2}\right)^{3/2} \exp(-Q/kT).$$

  a ratio of number densities

$$\frac{n(^8\mathrm{Be})}{n_\alpha} = 7 \times 10^{-9}.$$

This corresponds to an *equilibrium ^{8}Be concentration* of

$$n(^8\mathrm{Be}) = 10^{21}\ \mathrm{cm}^{-3}$$

during the flash (see Problem).

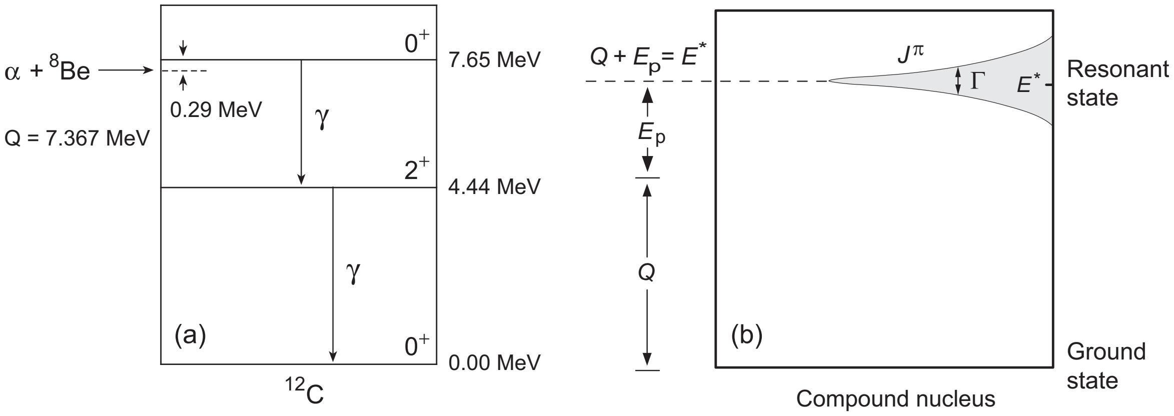


Figure 6.2: Nuclear energy levels in ^{12}C for the final steps of the triple-α reaction. Levels are labeled by J^π and energy relative to the ground state. The 0^+ state at 7.65 MeV is the Hoyle resonance.

6.3.2 Formation of the Excited State in Carbon-12

The next step of the triple-α process,

$$\alpha + {}^{8}\mathrm{Be} \rightarrow {}^{12}\mathrm{C} + \gamma$$

- has $Q = 7.367\,\mathrm{MeV}$ and
- proceeds through an angular momentum $J = 0$ resonance in ^{12}C at an excitation energy relative to the ^{12}C ground-state of 7.654 MeV, as illustrated in Fig. 6.2.

Hoyle resonance: The existence of this *Hoyle resonance* was predicted by Hoyle to explain the energy production in red giant stars.

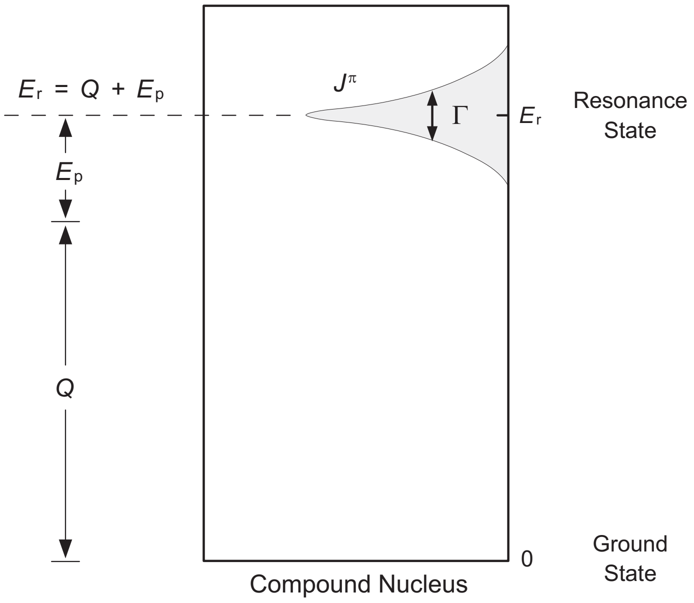


Figure 6.3: Relationship of Q-value, resonance energy E_r, and center of mass energy E_p when an isolated resonance is maximally excited.

- Population of the Hoyle state is optimized if the center of mass energy plus the Q-value equals the resonance energy relative to the ground state of ^{12}C (Fig. 6.3).
- Once this excited state is formed, the dominant reaction is a rapid decay back to $\alpha + {}^8$Be.
- However, a small fraction of the time the ground state of ^{12}C may instead be formed by two γ-ray decays (Fig. 6.2).
- If nuclear statistical equilibrium is assumed,

$$\frac{n\left({}^{12}\mathrm{C}^*\right)}{n_\alpha^3} = 3^{3/2}\left(\frac{h^2}{2\pi m_\alpha kT}\right)^3 \exp[(3m_\alpha - m_{12}^*)c^2/kT],$$

where m_{12}^* is the ^{12}C mass in the excited state (Problem).

6.3.3 Formation of the Ground State in Carbon-12

- Preceding considerations imply a *dynamical equilibrium*

$$^{4}\mathrm{He} + {}^{4}\mathrm{He} + {}^{4}\mathrm{He} \longleftrightarrow {}^{4}\mathrm{He} + {}^{8}\mathrm{Be} \longleftrightarrow {}^{12}\mathrm{C}^{*}.$$

- This produces an *equilibrium population of* $^{12}C^{*}$, almost all of which decays back to $^{4}\mathrm{He} + {}^{8}\mathrm{Be}$.
- However, the excited state of $^{12}\mathrm{C}$ can *decay electromagnetically to its ground state* with a mean life of

$$\tau\left({}^{12}\mathrm{C}^{*} \to {}^{12}\mathrm{C(gs)}\right) = 1.8 \times 10^{-16}\,\mathrm{s},$$

- This implies that *one in about every 2500 excited carbon nuclei that are produced* decay to the stable ground state.
- This decay probability is *very small*.
- Thus it *does not influence the equilibrium* appreciably and we may represent the triple-α process schematically as

$$^{4}\mathrm{He} + {}^{4}\mathrm{He} + {}^{4}\mathrm{He} \leftrightarrow {}^{4}\mathrm{He} + {}^{8}\mathrm{Be} \leftrightarrow {}^{12}\mathrm{C}^{*} \longrightarrow {}^{12}\mathrm{C(gs)}.$$

  - where left-right arrows indicate nuclear *statistical equilibrium* and
  - the one-way arrow indicates a *leakage from the equilibrium* that is a small perturbation;
  - since it is small, *it does not disturb the equilibrium* significantly.

- Thus in the approximate equilibrium

  $${}^{4}\mathrm{He} + {}^{4}\mathrm{He} + {}^{4}\mathrm{He} \ \leftrightarrow\ {}^{4}\mathrm{He} + {}^{8}\mathrm{Be} \ \leftrightarrow\ {}^{12}\mathrm{C}^{*} \ \longrightarrow\ {}^{12}\mathrm{C(gs)},$$

  the production rate for ^{12}C in its ground state is the product of

  - the *equilibrium ${}^{12}C^{*}$ population* and
  - the *decay rate to the ground state:*

  $$\begin{aligned}\frac{dn\left({}^{12}\mathrm{C}\right)}{dt} &= n({}^{12}\mathrm{C}^{*}) \times (\text{Decay rate } {}^{12}\mathrm{C}^{*} \to {}^{12}\mathrm{C(gs)}) \\ &= \frac{n\left({}^{12}\mathrm{C}^{*}\right)}{\tau\left({}^{12}\mathrm{C}^{*} \to {}^{12}\mathrm{C(gs)}\right)} \\ &= \frac{n_{\alpha}^{3}}{\tau\left({}^{12}\mathrm{C}^{*} \to {}^{12}\mathrm{C(gs)}\right)} 3^{3/2}\left(\frac{h^2}{2\pi m_{\alpha} kT}\right)^{3} \\ &\quad \times \ \exp[-(m_{12}^{*} - 3m_{\alpha})c^2/kT],\end{aligned}$$

  where we have used

  $$\frac{n\left({}^{12}\mathrm{C}^{*}\right)}{n_{\alpha}^{3}} = 3^{3/2}\left(\frac{h^2}{2\pi m_{\alpha} kT}\right)^{3} \exp[(3m_{\alpha} - m_{12}^{*})c^2/kT].$$

From the result

$$\frac{dn\left({}^{12}\mathrm{C}\right)}{dt} = \frac{n_\alpha^3}{\tau\left({}^{12}\mathrm{C}^* \to {}^{12}\mathrm{C(gs)}\right)} 3^{3/2} \left(\frac{h^2}{2\pi m_\alpha kT}\right)^3$$
$$\times \exp[-(m_{12}^* - 3m_\alpha)c^2/kT],$$

we see that the *rate of carbon production* depends on

1. The *number density* of α-particles n_α and *temperature* T.
2. An *activation energy* given by

   $$(m_{12}^* - 3m_\alpha)\, c^2 = 0.3795\,\mathrm{MeV}$$

   that must be borrowed to create the ${}^{12}\mathrm{C}^*$ intermediate state.
3. The *mean life for the decay ${}^{12}C^* \to {}^{12}C(gs)$*, which is

   $$\tau\left({}^{12}\mathrm{C}^* \to {}^{12}\mathrm{C(gs)}\right) = 1.8 \times 10^{-16}\,\mathrm{s}.$$

Table 6.2: Parameters for the triple-α reaction*

T (K)	kT (keV)	q/kT	$\exp(-q/kT)$
5×10^7	4.309	88.08	5.6×10^{-39}
1×10^8	8.617	44.04	7.5×10^{-20}
2×10^8	17.234	22.02	2.7×10^{-10}

*The activation energy is $q \equiv (m^*_{12} - 3m_\alpha)\,c^2$

The *strong temperature dependence* for the triple-α reaction results primarily from the *exponential Boltzmann factor* in

$$\frac{dn\,(^{12}\mathrm{C})}{dt} = \frac{n_\alpha^3}{\tau\,(^{12}\mathrm{C}^* \to {}^{12}\mathrm{C(gs)})}\,3^{3/2}\left(\frac{h^2}{2\pi m_\alpha kT}\right)^3 \times \exp[-(m^*_{12} - 3m_\alpha)c^2/kT],$$

because at helium burning temperatures

- the average thermal energy kT is typically much less than the activation energy of $Q = -379.5\,\mathrm{keV}$.
- This is illustrated in Table 6.2.

From Table 6.2 we see that

- *doubling the temperature* near 10^8 K
- changes the Boltzmann factor by *10–20 orders of magnitude*.

6.3.4 Energy Production in the Triple-α Reaction

The total energy released in the triple-α reaction is

$$Q = 7.275\,\mathrm{MeV}$$

and the energy production rate is given by

$$\varepsilon_{3\alpha} = \frac{5.1\times 10^8 \rho^2 Y^3}{T_9^3} \exp(-4.4027/T_9)\,\mathrm{erg\,g^{-1}s^{-1}},$$

where Y is the helium abundance. From

$$\lambda = \left(\frac{\partial \ln \varepsilon}{\partial \ln \rho}\right)_T \qquad \nu = \left(\frac{\partial \ln \varepsilon}{\partial \ln T}\right)_\rho$$

- This implies *density and temperature exponents*

$$\lambda_{3\alpha} = 2 \qquad \nu_{3\alpha} = \frac{4.4}{T_9} - 3,$$

- The quadratic dependence on the density occurs because the *reaction is effectively 3-body*.
- For $T_8 = 1$, we obtain a temperature exponent

$$\nu_{3\alpha} \simeq 40$$

→ a helium core is a *very explosive fuel*.

If a fuel has a large temperature exponent, the rate of burning can increase enormously if the temperature increases only a little. This greatly increases the probability that burning becomes explosive.

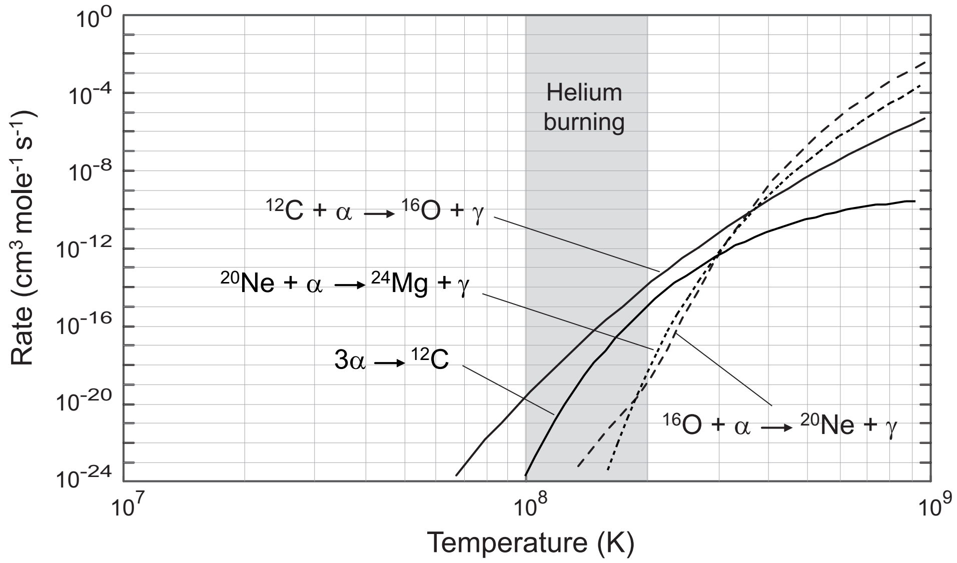


Figure 6.4: Triple-α and radiative α-capture rates important in helium burning. Vertical gray band indicates the temperature range for helium burning.

6.4 Burning of Carbon to Oxygen and Neon

- Once carbon has been formed by triple-α, oxygen can be produced by

$$^4\text{He} + {}^{12}\text{C} \to {}^{16}\text{O} + \gamma.$$

- which has *no resonances* near the Gamow window.
- The rate is
  - *slow* and is
  - experimentally *rather uncertain*.

  The currently accepted rate is shown in Fig. 6.4.
- The uncertainty has consequences because this rate *determines the ratio of C to O production* in stars.

- The rate for

  $$^{4}\mathrm{He} + {}^{12}\mathrm{C} \rightarrow {}^{16}\mathrm{O} + \gamma.$$

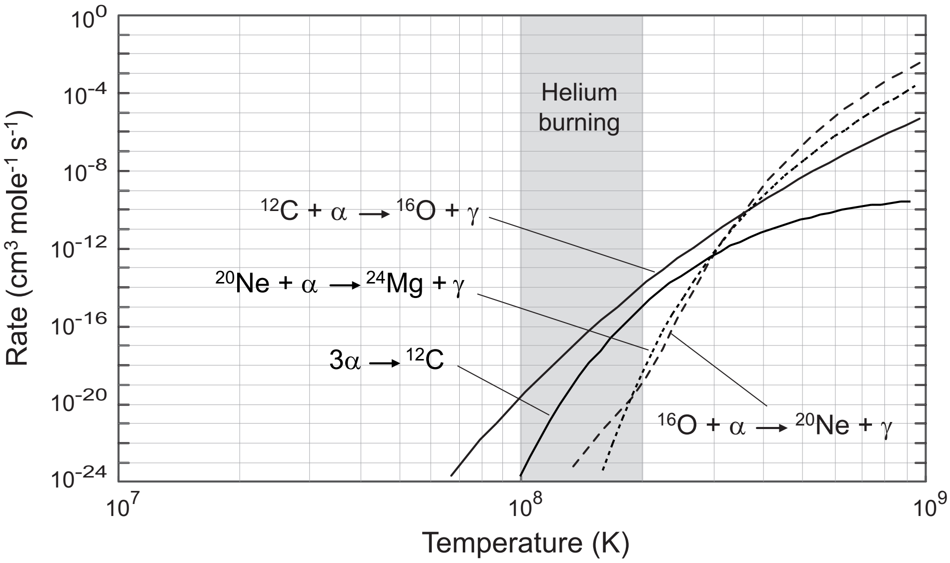
  impacts the *abundance of C and O in the Universe*, but also
- the carbon–oxygen ratio in stellar cores can have *large influence on late stellar evolution*.

For example,

- the *composition of white dwarfs* and
- the *composition of cores of massive stars* late in their lives

depend critically on this rate, so it can have a large impact on

- *how stars die* and
- *what is left behind* when they do.



- Once oxygen is produced by

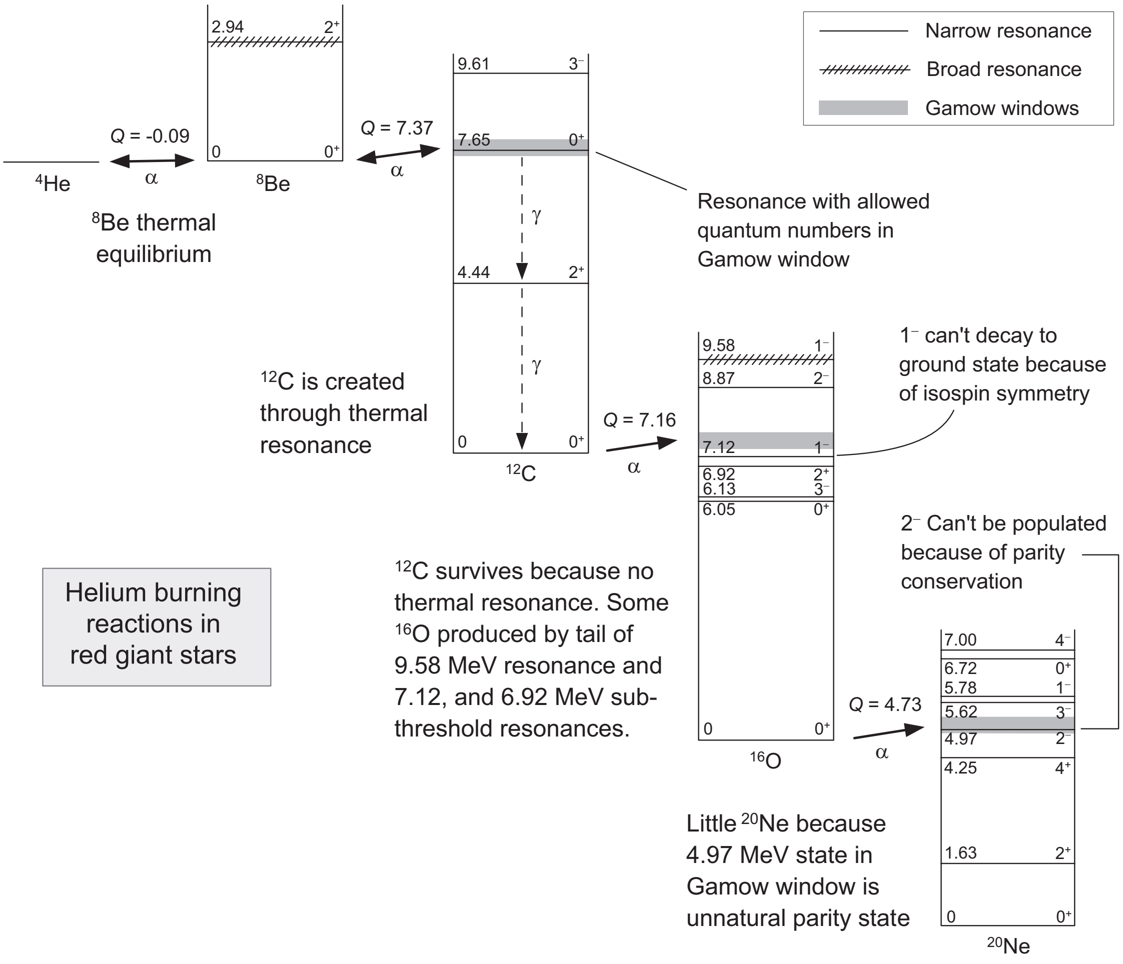
$$^4\text{He} + {}^{12}\text{C} \rightarrow {}^{16}\text{O} + \gamma,$$

neon can be formed by (rate plotted in figure above)

$$^4\text{He} + {}^{16}\text{O} \rightarrow {}^{20}\text{Ne} + \gamma.$$

- The reaction is *slow* because
  - it is *nonresonant* and
  - it has a *large Coulomb barrier*.
- Thus, *little neon is produced* during helium burning and
- The primary residue of helium burning is a C–O core:
  - The carbon is produced by the triple-α sequence and
  - the oxygen by radiative capture on the carbon,

with the C–O ratio depending on an uncertain rate.



8Be thermal equilibrium

12C is created through thermal resonance

Resonance with allowed quantum numbers in Gamow window

1− can't decay to ground state because of isospin symmetry

2− Can't be populated because of parity conservation

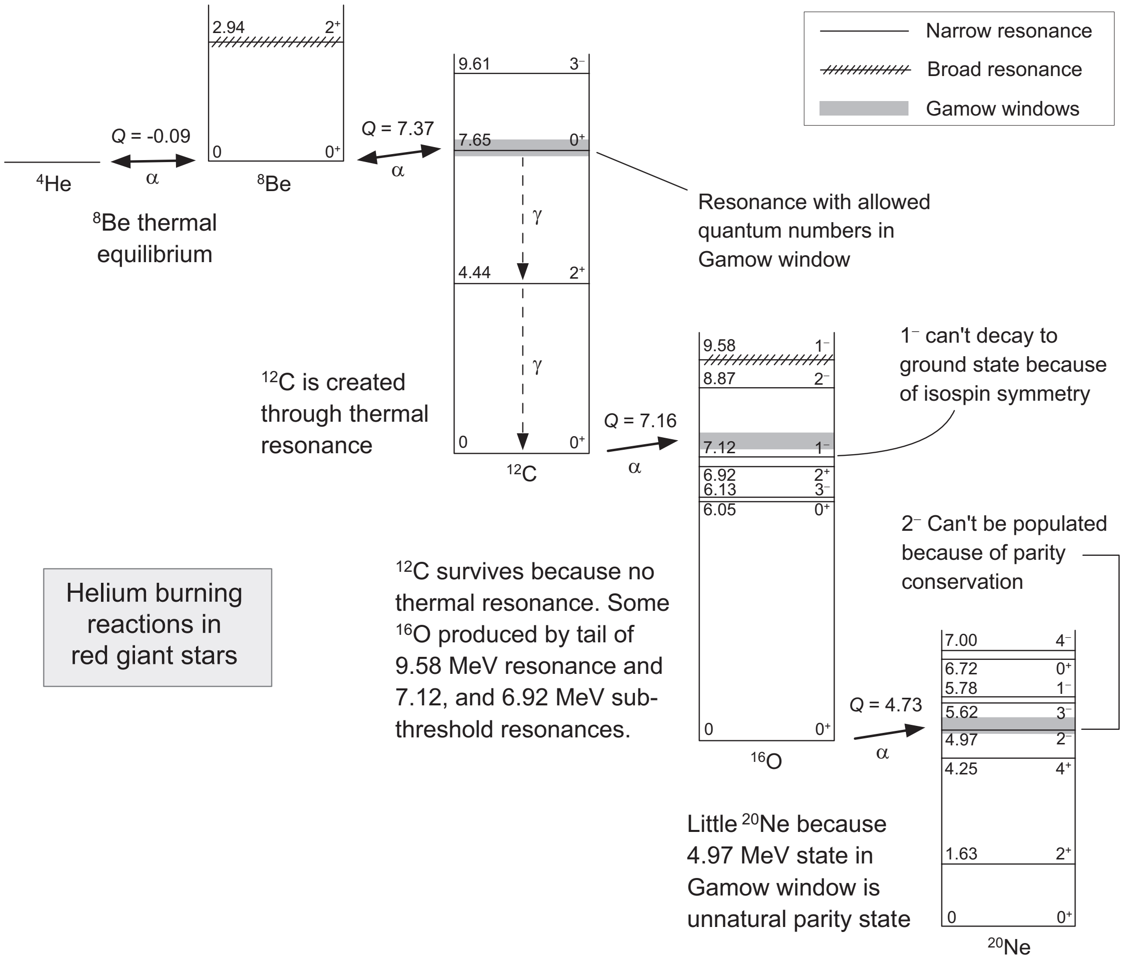
12C survives because no thermal resonance. Some 16O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.

Little 20Ne because 4.97 MeV state in Gamow window is unnatural parity state

Figure 6.5: Overview of helium burning.

6.5 The Outcome of Helium Burning

The outcome of helium burning is summarized in Fig. 6.5. This outcome is a remarkable example of how fundamentally different our Universe would be if just a few seemingly boring details of nuclear physics were slightly different.



^{8}Be thermal equilibrium

Resonance with allowed quantum numbers in Gamow window

^{12}C is created through thermal resonance

1$^-$ can't decay to ground state because of isospin symmetry

Helium burning reactions in red giant stars

^{12}C survives because no thermal resonance. Some ^{16}O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.

2$^-$ Can't be populated because of parity conservation

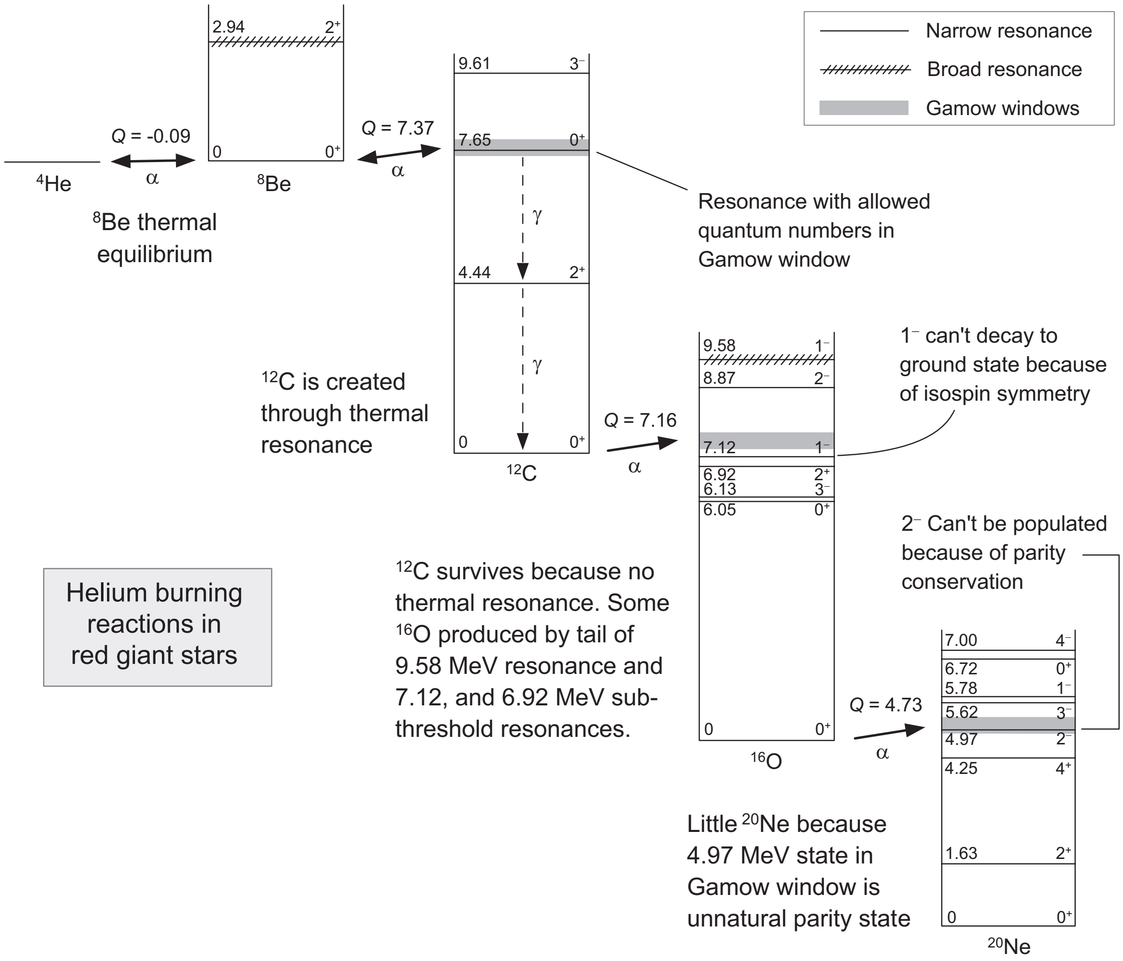
Little ^{20}Ne because 4.97 MeV state in Gamow window is unnatural parity state

- The ratio of C to O is determined by competition between
  - the C-producing triple-α reaction and
  - the C-depleting, O-producing capture reaction

$$^4\mathrm{He} + {}^{12}\mathrm{C} \rightarrow {}^{16}\mathrm{O} + \gamma.$$

- Further, that much C or O exists at all is dependent on the slowness of the Ne-producing reaction

$$^4\mathrm{He} + {}^{16}\mathrm{O} \rightarrow {}^{20}\mathrm{Ne} + \gamma.$$

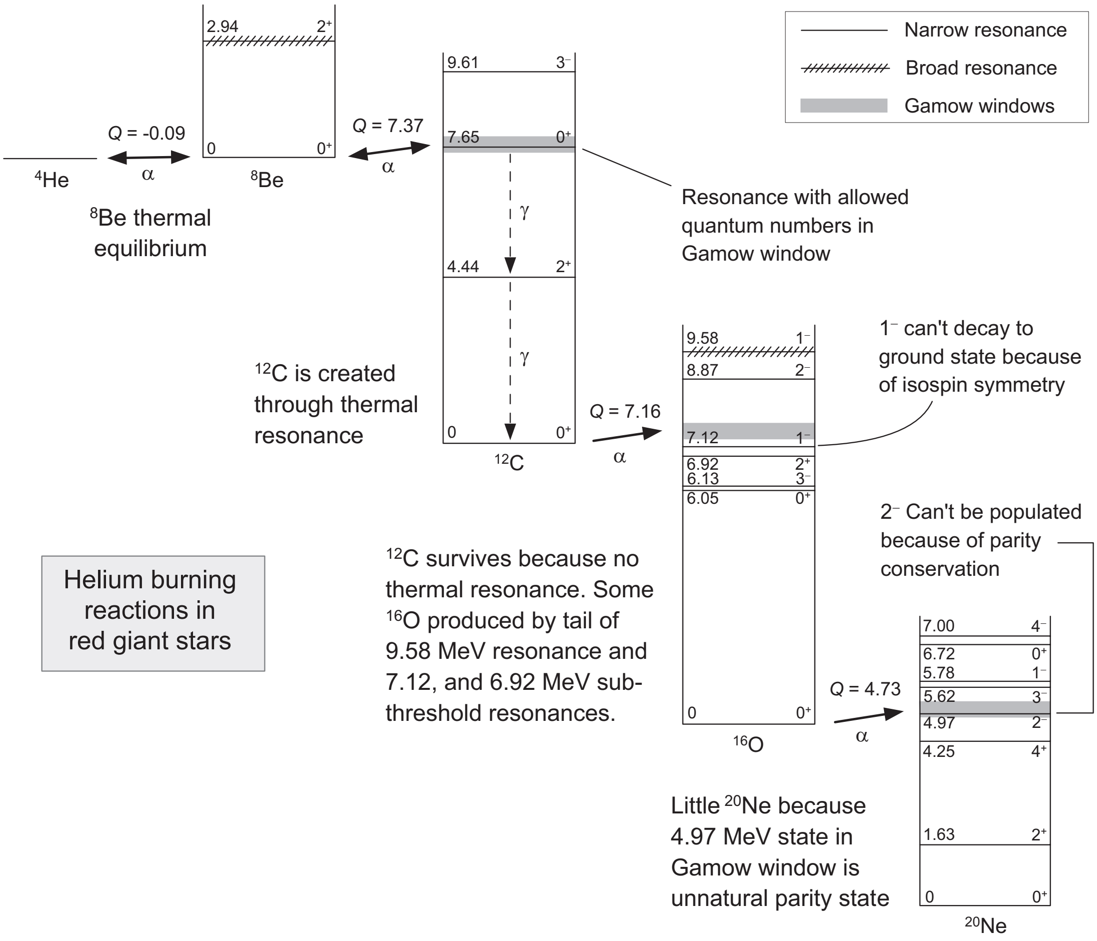

Narrow resonance
Broad resonance
Gamow windows
^{4}He
Q = -0.09
α
2.94 2⁺
0 0⁺
^{8}Be
^{8}Be thermal equilibrium
Q = 7.37
α
9.61 3⁻
7.65 0⁺
γ
4.44 2⁺
γ
0 0⁺
^{12}C
Resonance with allowed quantum numbers in Gamow window
^{12}C is created through thermal resonance
Q = 7.16
α
9.58 1⁻
8.87 2⁻
7.12 1⁻
6.92 2⁺
6.13 3⁻
6.05 0⁺
0 0⁺
^{16}O
1⁻ can't decay to ground state because of isospin symmetry
2⁻ Can't be populated because of parity conservation
Helium burning reactions in red giant stars
^{12}C survives because no thermal resonance. Some ^{16}O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.
Q = 4.73
α
7.00 4⁻
6.72 0⁺
5.78 1⁻
5.62 3⁻
4.97 2⁻
4.25 4⁺
1.63 2⁺
0 0⁺
^{20}Ne
Little ^{20}Ne because 4.97 MeV state in Gamow window is unnatural parity state

- If, contrary to fact, a resonance existed near the fusion window for

$$^4\mathrm{He} + {}^{12}\mathrm{C} \rightarrow {}^{16}\mathrm{O} + \gamma,$$

  - the corresponding *rate would be large* and
  - almost all carbon produced by triple-α would be *converted rapidly to oxygen*,

leaving *little carbon in the Universe*.

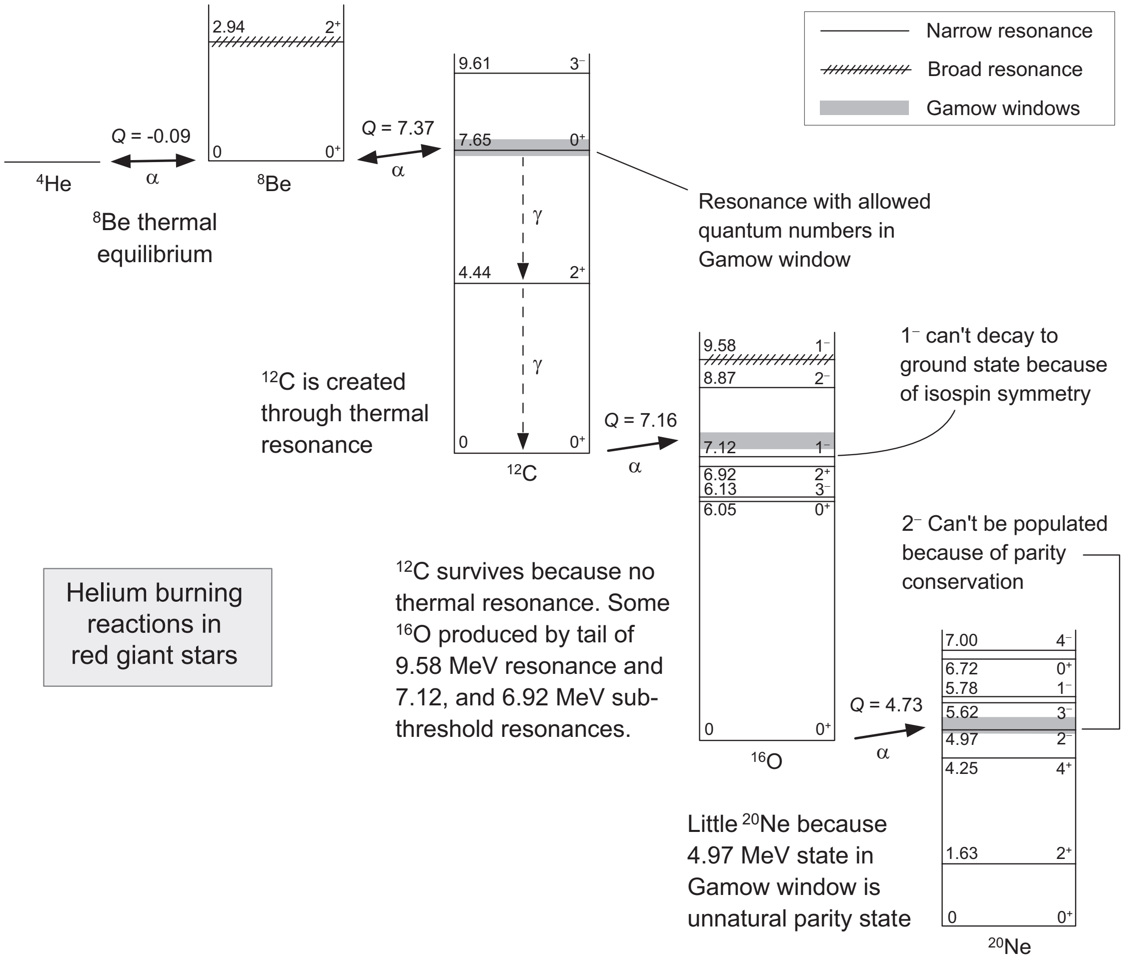

Narrow resonance
Broad resonance
Gamow windows
2.94 2+
0 0+
Q = -0.09
α
4He
8Be
8Be thermal equilibrium
Q = 7.37
α
9.61 3−
7.65 0+
γ
4.44 2+
γ
0 0+
12C
Resonance with allowed quantum numbers in Gamow window
12C is created through thermal resonance
Q = 7.16
α
9.58 1−
8.87 2−
7.12 1−
6.92 2+
6.13 3−
6.05 0+
0 0+
16O
1− can't decay to ground state because of isospin symmetry
2− Can't be populated because of parity conservation
Helium burning reactions in red giant stars
12C survives because no thermal resonance. Some 16O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.
Q = 4.73
α
7.00 4−
6.72 0+
5.78 1−
5.62 3−
4.97 2−
4.25 4+
1.63 2+
0 0+
20Ne
Little 20Ne because 4.97 MeV state in Gamow window is unnatural parity state

- A similar fate would follow if the Hoyle resonance at 7.65 MeV were *a little higher in energy*, greatly slowing triple-α by virtue of the Boltzmann factor in

$$\frac{dn\left({}^{12}\mathrm{C}^*\right)}{dt} = \frac{n_\alpha^3}{\tau\left({}^{12}\mathrm{C}^* \to {}^{12}\mathrm{C(gs)}\right)}\, 3^{3/2}\left(\frac{h^2}{2\pi m_\alpha kT}\right)^3$$
$$\times\ \exp[-(m_{12}^* - 3m_\alpha)c^2/kT],$$

and any C produced would be converted rapidly to O by

$$^4\mathrm{He} + {}^{12}\mathrm{C} \to {}^{16}\mathrm{O} + \gamma.$$



^{8}Be thermal equilibrium

Resonance with allowed quantum numbers in Gamow window

^{12}C is created through thermal resonance

1$^-$ can't decay to ground state because of isospin symmetry

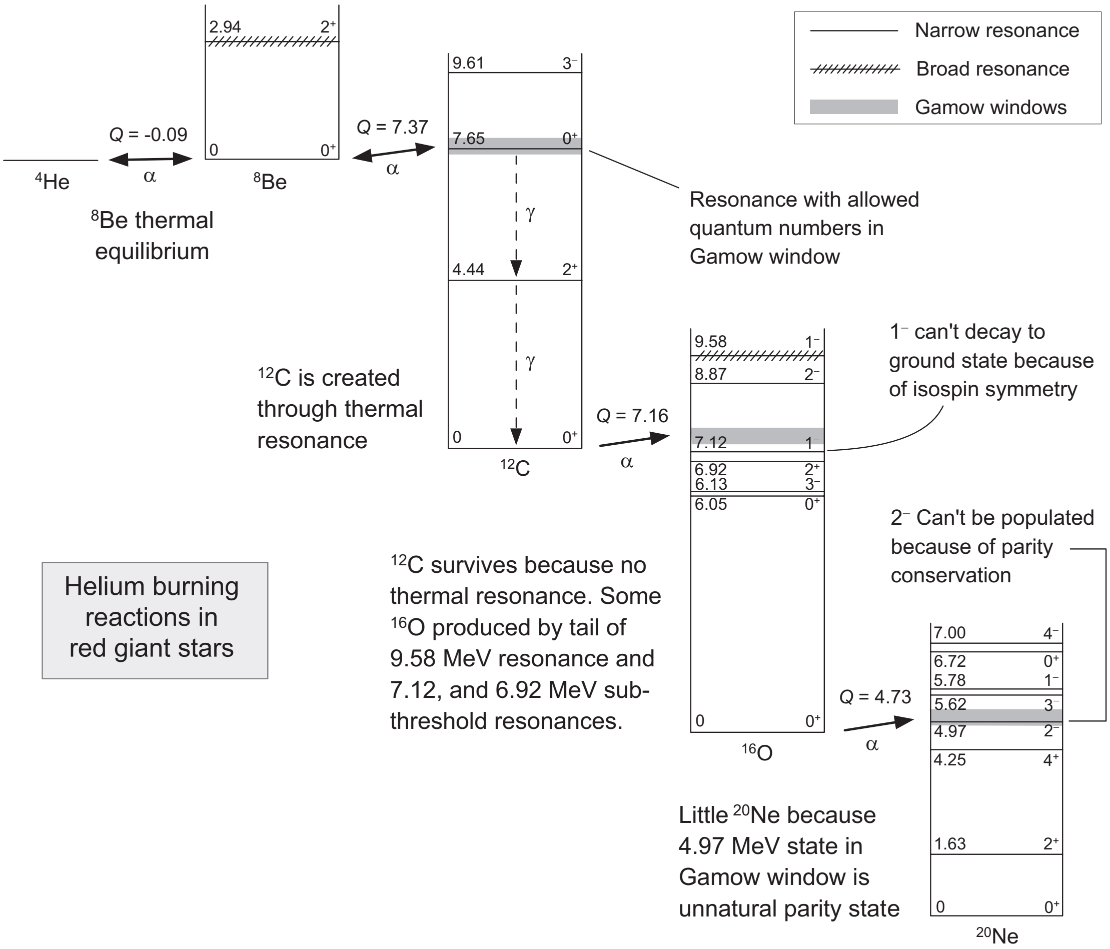
Helium burning reactions in red giant stars

^{12}C survives because no thermal resonance. Some ^{16}O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.

2$^-$ Can't be populated because of parity conservation

Little ^{20}Ne because 4.97 MeV state in Gamow window is unnatural parity state

- On the other hand, if the resonance at 7.65 MeV in ^{12}C *did not exist*,
  - *Triple-α would not work at all* in red giant stars and
  - there would be *little carbon or oxygen* in the Universe.



^{8}Be thermal equilibrium

^{12}C is created through thermal resonance

Resonance with allowed quantum numbers in Gamow window

1− can't decay to ground state because of isospin symmetry

2− Can't be populated because of parity conservation

Helium burning reactions in red giant stars

^{12}C survives because no thermal resonance. Some ^{16}O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.

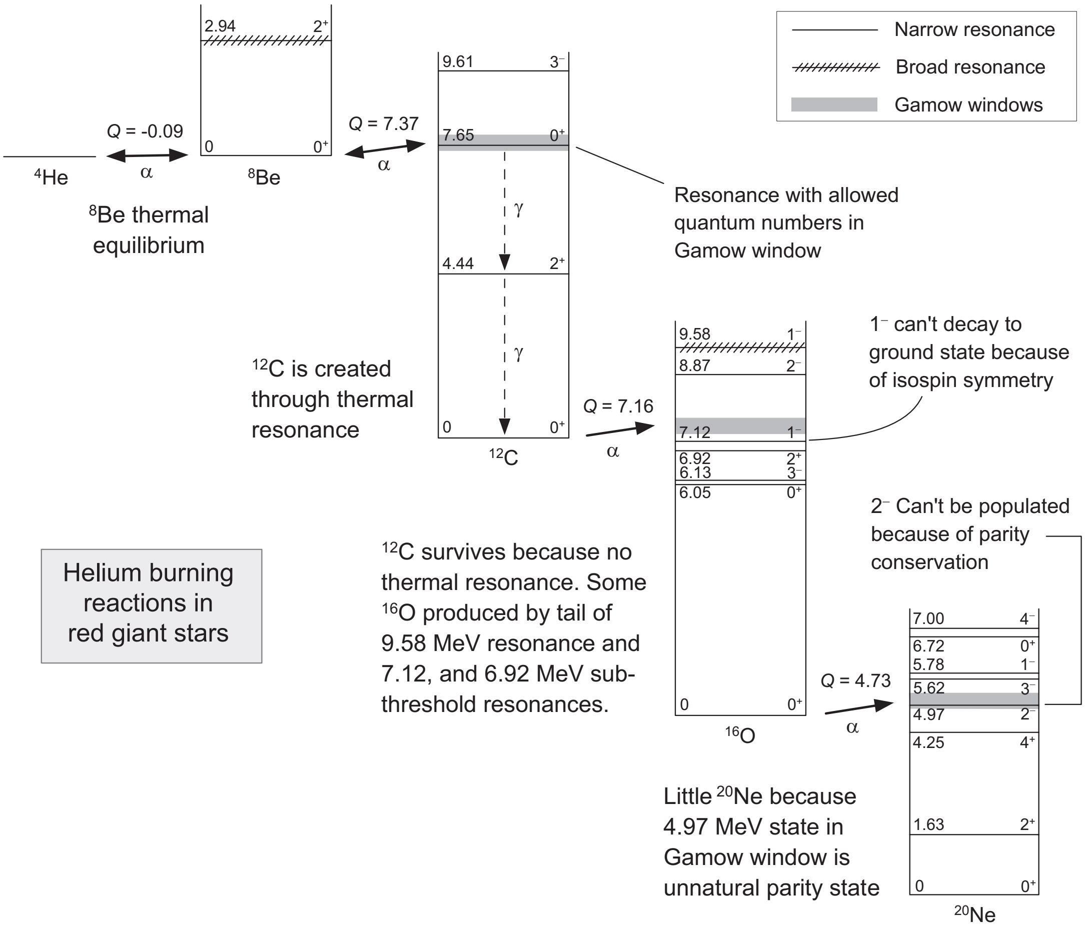
Little ^{20}Ne because 4.97 MeV state in Gamow window is unnatural parity state

- Finally, if

$$^4\text{He} + {}^{16}\text{O} \rightarrow {}^{20}\text{Ne} + \gamma$$

were *resonant*—which it would be if the *parity of a single excited state in neon* were positive instead of negative

  - Most of the C and O produced by helium burning would be *transformed by this reaction to Ne*.
  - Neon is a *noble gas* and therefore *chemically inert*,
  - This contrasts with the rich chemistry of carbon that *makes biology possible* in the actual Universe.


Narrow resonance
Broad resonance
Gamow windows
2.94 2+
0 0+
^{4}He
Q = -0.09
α
^{8}Be
^{8}Be thermal equilibrium
Q = 7.37
α
9.61 3−
7.65 0+
γ
4.44 2+
γ
0 0+
^{12}C
Resonance with allowed quantum numbers in Gamow window
^{12}C is created through thermal resonance
Q = 7.16
α
9.58 1−
8.87 2−
7.12 1−
6.92 2+
6.13 3−
6.05 0+
0 0+
^{16}O
1− can't decay to ground state because of isospin symmetry
2− Can't be populated because of parity conservation
Helium burning reactions in red giant stars
^{12}C survives because no thermal resonance. Some ^{16}O produced by tail of 9.58 MeV resonance and 7.12, and 6.92 MeV sub-threshold resonances.
Q = 4.73
α
7.00 4−
6.72 0+
5.78 1−
5.62 3−
4.97 2−
4.25 4+
1.63 2+
0 0+
^{20}Ne
Little ^{20}Ne because 4.97 MeV state in Gamow window is unnatural parity state

Thus, our very existence appears to depends on the *parity* of obscure nuclear states in atoms that have nothing whatsoever to do with the chemistry of life!

The Anthropic Principle and Helium Burning:

Many would argue, based on the

- observed *diversity of life* on Earth and
- *how quickly it arose* after formation of the planet,

that life in the Universe is inevitable.

- But this point of view assumes the existence of the chemicals on which life (as we know it) is built.
- Our discussion suggests that the existence of the building blocks of life depends on arcane facts on the MeV scale (nuclear physics)
- that have nothing to do with the physics of eV scales (chemistry) that governs life.
- The very possibility of biochemistry may be an *accident of physical parameter values* in our Universe.
- Such considerations lie at the basis of the (simplest) *anthropic principle:*

> The Universe has just the right value of constants and just the right detailed physics required for life because, if it didn't, there would be no life in the Universe and thus no one to ask the question.

- Is this line of thinking is even scientific (*is it testable*)?

6.6 Advanced Burning Stages

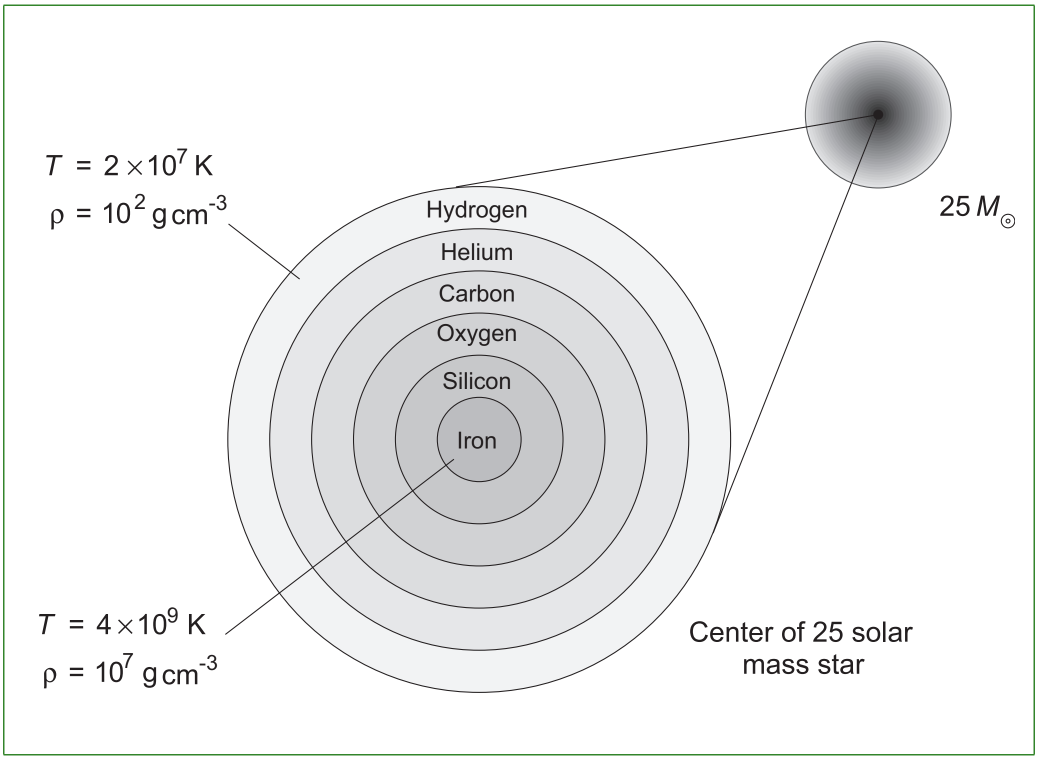
If a star is massive enough,

- more *advanced burnings* are possible by virtue of the
- *high temperatures* and *high densities* that result as the core contracts after exhausting its fuel.

Typical burning stages in massive stars and their characteristics are illustrated in the following table and figure.

Table 6.3: Burning stages in massive stars (Woosley)

Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day



Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day

Carbon burning: Carbon burns at a temperature of

$$T \sim 5\times10^8\,\mathrm{K}$$

and a density of

$$\rho \sim 3\times10^6\,\mathrm{g\,cm^{-3}},$$

primarily through the reactions

$$^{12}\mathrm{C} + {}^{12}\mathrm{C} \longrightarrow {}^{20}\mathrm{Ne} + \alpha$$
$$^{12}\mathrm{C} + {}^{12}\mathrm{C} \longrightarrow {}^{23}\mathrm{Na} + p$$
$$^{12}\mathrm{C} + {}^{12}\mathrm{C} \longrightarrow {}^{23}\mathrm{Mg} + p$$

As indicated in the Table above, such reactions are possible for stars having masses larger than about $4M_\odot$.

- Burning stages beyond that of carbon require conditions that are realized only for stars having $M \gtrsim 8M_\odot$ or so.
- At the required temperatures, a new feature comes into play because *the most energetic photons can disrupt the nuclei produced in preceding burning stages.*

Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day

Neon burning: At $T \sim 10^9$ K, neon can burn by a two-step sequence.

- First, a neon nucleus is *photodisintegrated* by a high-energy photon

$$\gamma + {}^{20}\mathrm{Ne} \longrightarrow {}^{16}\mathrm{O} + \alpha,$$

  which become *more plentiful at high temperature* since the average photon energy is $\sim kT$.

- Then the alpha-particle produced in this step can initiate a *radiative capture reaction*

$$\alpha + {}^{20}\mathrm{Ne} \rightarrow {}^{24}\mathrm{Mg} + \gamma.$$

This burning sequence produces a *core of* ${}^{16}O$ *and* ${}^{24}Mg$.

Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day

Oxygen burning: At a temperature of 2×10^9 K,

- Oxygen can fuse through the reaction

$$^{16}\mathrm{O} + {}^{16}\mathrm{O} \longrightarrow {}^{28}\mathrm{Si} + \alpha$$

- The silicon thus produced can react only at temperatures where photodissociation reactions begin to play a dominating role.

Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day

Silicon burning: At $T \sim 3 \times 10^9$ K, silicon may be burned to heavier elements.

- At these temperatures the *photons are quite energetic* and
- those in the high-energy tail of the Maxwell–Boltzmann distribution can readily *photodissociate nuclei*.
- A network of photodisintegration and capture reactions in approximate *nuclear statistical equilibrium* develops and
- the population in this network evolves preferentially to those isotopes that have the *largest binding energies*.
- From the binding energy curve the *most stable nuclei are in the iron group*.

Silicon burning carried to completion under equilibrium conditions *produces iron-group nuclei*.

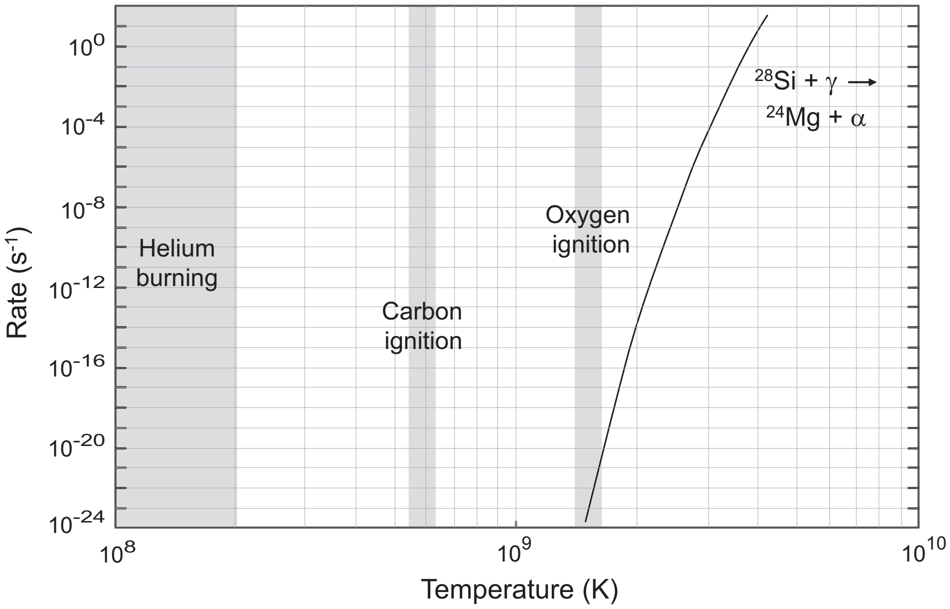


Figure 6.6: Temperature dependence of the highly-endothermic, rate-controlling initial step in silicon burning. For reference, the typical range of temperatures corresponding to helium burning and for carbon and oxygen ignition are indicated. Silicon burning requires temperatures more than an order of magnitude larger than for helium burning, and exhibits an extremely strong temperature dependence.

- The initial step in Si burning is a photodisintegration like

$$\gamma + {}^{28}\text{Si} \longrightarrow {}^{24}\text{Mg} + \alpha,$$

  which requires a photon energy of 9.98 MeV or greater.

- The temperature dependence is illustrated in Fig. 6.6.

- From this plot we infer that

  - silicon burning depends strongly on temperature, and
  - it requires temperatures more than an order of magnitude larger than for helium burning.

- The α particles thus liberated can now initiate radiative capture reactions on seed isotopes in the gas.
- A representative sequence is

$$\begin{gathered}\alpha + {}^{28}\mathrm{Si} \longleftrightarrow {}^{32}\mathrm{S} + \gamma \\ \alpha + {}^{32}\mathrm{S} \longleftrightarrow {}^{36}\mathrm{Ar} + \gamma \\ \vdots \\ \alpha + {}^{52}\mathrm{Fe} \longleftrightarrow {}^{56}\mathrm{Ni} + \gamma\end{gathered}$$

- The reactions in this series are typically in equilibrium or quasiequilibrium, and
- they are much faster than the initial photodisintegration.

Thus the *photodisintegration of silicon is the rate-controlling step* in silicon burning.

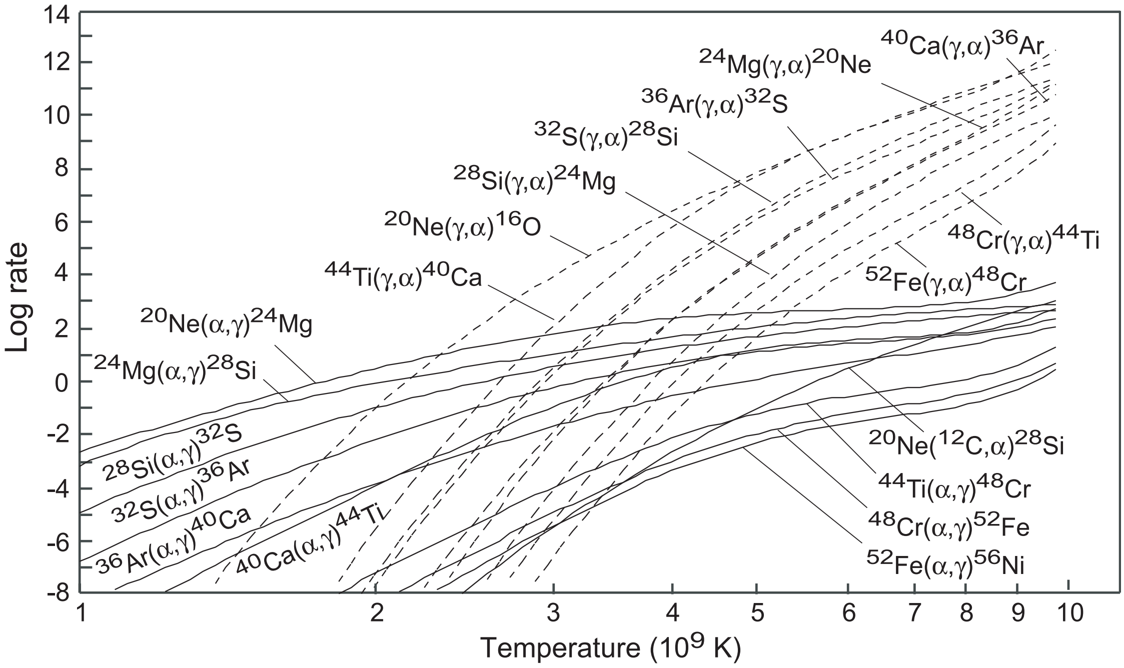


Figure 6.7: Some rates for competing capture reactions $A(\alpha,\gamma)B$ and photodisintegration reactions $A(\gamma,\alpha)B$ that are important for silicon burning. Photodisintegration rates are in units of s^{-1} and α-capture rates are in units of $\mathrm{cm}^3\,\mathrm{mol}^{-1}\,\mathrm{s}^{-1}$.

- The rates for some competing capture and photodisintegration reactions in Si burning are illustrated in Fig. 6.7.
- Note the steep T dependence of the photodisintegrations.

For high T and α-particle abundance, many photodisintegration rates become comparable to the rates for their inverse capture reactions somewhere in the range $T \sim 10^9 - 10^{10}$ K.

- The iron group nuclei are the most stable in the Universe.
- Thus silicon burning represents the last stage by which fusion and radiative capture reactions can build heavier elements under equilibrium conditions.
- One might think that we could make still heavier elements by increasing the temperature.
- Then the required extra energy for fusion presented by higher Coulomb barriers could be provided by the kinetic energy of the gas.
- But this becomes self-defeating in equilibrium:

  1. the higher temperatures will also lead to increased photodissociation.
  2. Thus iron-group nuclides are still the equilibrium product.

In subsequent chapters we will address the issue of other mechanisms by which stars can produce the elements heavier than iron.

Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day

6.7 Timescales for Advanced Burning

As is apparent from the table above,

- the timescales for advanced burning are greatly compressed relative to earlier burning stages.
- These differences are particularly striking for massive stars, which rush through all stages at breakneck speed.

For example, the $25M_\odot$ example used for the above table

- Takes about *10 million years* to advance through its hydrogen and helium burning phases,
- completes its burning of oxygen in only *six months*, and
- transforms its newly-minted silicon into iron group nuclei in *a single day*.

Nuclear fuel	Nuclear products	Ignition temperature	Minimum main sequence mass	Period in $25M_\odot$ star
H	He	4×10^6 K	$0.1M_\odot$	7×10^6 years
He	C, O	1.2×10^8 K	$0.4M_\odot$	5×10^5 years
C	Ne, Na, Mg, O	6×10^8 K	$4M_\odot$	600 years
Ne	O, Mg	1.2×10^9 K	$\sim 8M_\odot$	1 years
O	Si, S, P	1.5×10^9 K	$\sim 8M_\odot$	~ 0.5 years
Si	Ni–Fe	2.7×10^9 K	$\sim 8M_\odot$	~ 1 day

These timescales are set by

- the *amount of fuel available*,
- the *energy per reaction* derived from burning the fuel, and
- the *rate of energy loss* from the star, which ultimately governs the burning rate.
- This last factor is particularly important because
- *energy losses are large* when the reaction must run at *high temperature*.

Each factor separately shortens the timescale for advanced burning; taken together they make the timescales for the most advanced burning *almost instantaneous on the scale set by the hydrogen burning*.

An Analogy: To get a perspective on how short the advanced burning timescape is, imagine the lifetime of the $25M_\odot$ star to be *compressed into a single year*. Then

- the *hydrogen* fuel would be gone by about *December 7* of that year,
- the *helium* would burn over the *next 24 days*,
- the *carbon* would burn in the *42 minutes before midnight*,
- the *neon and oxygen* would burn in the *last seconds before midnight*, and
- the *silicon* would be converted to *iron* in the *last 1/100 second of the year* ,
- (with a quite impressive New Year's Eve *fireworks display* in the offing—see the later discussion of core-collapse supernovae).

Chapter 7

Energy Transport in Stars

- Most energy production in stars takes place in the deep interior where density and temperatures are high, but
- most electromagnetic energy that we see coming from stars is radiated from the photosphere, which is a very thin layer at the surface.

Thus, a fundamental issue in stellar astrophysics is *how energy is transported* from the interior to the surface.

7.1 Modes of Energy Transport

Energy transport in stars results from four general mechanisms:

1. *Conduction* because of thermal motion of electrons and ions,

2. *Radiative transport* by photons,

3. *Convection* of macroscopic packets of gas,

4. *Neutrino emission* from the core.

We may make a number of general statements about these modes of energy transport:

- Both
  - *conduction* and
  - *radiative transport*

  result from *random thermal motion* of constituent particles (electrons in the first case and photons in the second),
- *Convection* is a macroscopic or collective phenomenon.
- In normal stars *conduction* is negligible.
- However, *conduction* can be important in star containing degenerate matter (e.g., white dwarfs).
- *Radiative transport* usually dominates
  - unless the *temperature gradient* in the gravitational field exceeds a critical value.
  - If the *temperature gradient* becomes too steep, *convection* quickly becomes the most efficient means of energy transport.
- *Neutrino emission* is important for core cooling late in the life of *more massive stars*.

Neutrino emission differs from the other energy transport mechanisms in that

- it can operate only at extremely *high temperatures and densities*, and
- the neutrinos have *little interaction with the star* as they carry energy out of the core at essentially light speed.
- As a result, the first three modes of energy transport:
  - conduction,
  - radiation,
  - convection

  typically lead to *thermalization of the energy* (sharing of the energy among many particles).
- This *thermalized energy* is then eventually emitted as light of various wavelengths from the photosphere of the star.
- But the *energy of the neutrinos* almost always is carried away by the emitted neutrinos.

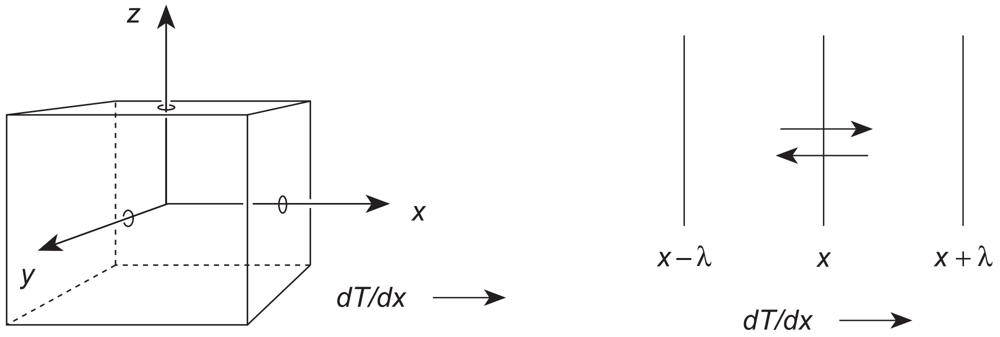
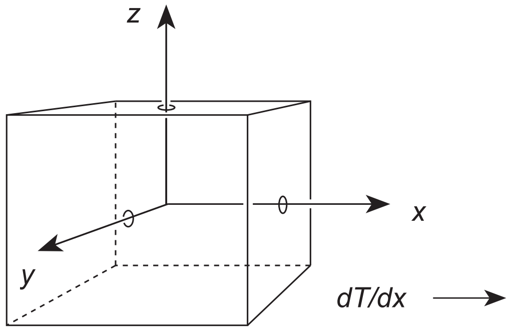


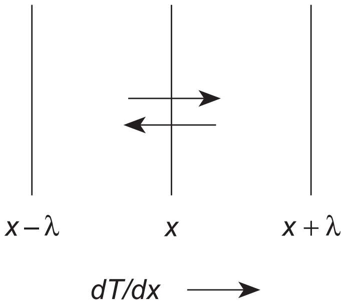
Figure 7.1: Diffusion of energy.

7.2 Diffusion of Energy

We begin with a discussion of how energy can be transported by random thermal motion (*diffusion*).

- Consider the volume enclosed by a small cube illustrated in Fig. 7.1.
- Introduce a *random velocity distribution* with a small *temperature gradient* in the x direction.
- On average, we may assume that at any instant approximately $\frac{1}{6}$ of the particles move in the positive x direction with mean velocity $\langle v \rangle$ and mean free path λ.
- Let $u(x)$ be the *thermal energy density*.
- Because of the *temperature gradient*, particles crossing a plane at x from left to right have a different thermal energy than those crossing from right to left (Fig. 7.1).
- Therefore, *energy is transported across the surface* by virtue of the *temperature gradient*.





- The rate of this transport is given by the current $j(x)$,

$$\begin{aligned} j(x) &\simeq \frac{1}{6}\langle v\rangle u(x-\lambda) - \frac{1}{6}\langle v\rangle u(x+\lambda) \\ &\simeq -\frac{1}{3}\langle v\rangle\lambda\frac{du}{dx} = -\frac{1}{3}\langle v\rangle\lambda\frac{du}{dT}\frac{dT}{dx} \\ &\simeq -\frac{1}{3}\langle v\rangle\lambda C\frac{dT}{dx}, \end{aligned}$$

  where the *heat capacity per unit volume* is $C = du/dT$.

- Therefore, the *current across the surface* may be written

$$j(x) = -\underbrace{\frac{1}{3}\langle v\rangle\lambda C}_{\equiv K}\frac{dT}{dx} = -K\frac{dT}{dx} \qquad \text{(Ficke's Law)},$$

  where K is termed the *coefficient of thermal conductivity:*

$$K \equiv \frac{1}{3}\langle v\rangle\lambda C.$$

- This equation for the current is sometimes termed *Ficke's Law;* it is *characteristic of diffusive processes*.

Although we have obtained Ficke's law for diffusion,

$$j(x) = -K\frac{dT}{dx} \qquad K \equiv \frac{1}{3}\langle v\rangle \lambda C,$$

in a carelessly heuristic way, a more careful derivation gives essentially the same result.

7.3 Energy Transport by Conduction

Let's first consider heat transport by *random motion* of electrons and ions.

- For a *nonrelativistic ideal gas of electrons* the
  - internal energy density u_e,
  - heat capacity C_e, and
  - average velocity $\langle v_\mathrm{e} \rangle$

  are given by

$$u_\mathrm{e} = \frac{3}{2} n_\mathrm{e} kT \qquad C_\mathrm{e} = \frac{3}{2} n_\mathrm{e} k \qquad \langle v_\mathrm{e} \rangle = \sqrt{3kT/m_\mathrm{e}}.$$

- The *electron–electron collisions* are much less effective than *electron–ion collisions* in transferring energy.
- Thus the relevant mean free path $\lambda = 1/n\sigma$ is

$$\lambda_\mathrm{ei} = \frac{1}{n_\mathrm{i} \sigma_\mathrm{ei}},$$

  where
  - n_i is the *number density of ions* and
  - σ_ei is the *cross section for electron–ion collisions*.

7.3.1 Coefficient of Thermal Conduction

As a first crude estimate of the *electron–ion cross section* we may assume

$$\sigma_{\mathrm{ei}} \simeq \pi R^2,$$

where R is the separation between electron and ion where the potential energy is equal to the average kinetic energy in the gas (kT),

$$Ze^2/R \simeq kT \quad \rightarrow \quad R \simeq \frac{kT}{Ze^2}.$$

Thus, the cross section is approximately

$$\sigma_{\mathrm{ei}} = \pi R^2 = \pi \left(\frac{Ze^2}{kT} \right)^2,$$

and substitution of $\lambda = 1/n\sigma$ in

$$K \equiv \frac{1}{3} \langle v \rangle \lambda C$$

yields

$$K_{\mathrm{e}} = \frac{k}{2\pi} \left(\frac{n_{\mathrm{e}}}{n_{\mathrm{i}}} \right) \left(\frac{kT}{Ze^2} \right)^2 \sqrt{\frac{3kT}{m_{\mathrm{e}}}}.$$

The corresponding expression for ionic conduction is obtained by the exchanges $n_{\mathrm{e}} \leftrightarrow n_{\mathrm{i}}$ and $m_{\mathrm{e}} \leftrightarrow m_{\mathrm{i}}$, and we obtain

$$\frac{K_{\mathrm{e}}}{K_{\mathrm{i}}} = \frac{n_{\mathrm{e}}^2}{n_{\mathrm{i}}^2} \sqrt{\frac{m_{\mathrm{i}}}{m_{\mathrm{e}}}}.$$

As an estimate we may assume the gas to be *completely ionized*, so that $n_\mathrm{e} = Zn_\mathrm{i}$ and

$$\frac{K_\mathrm{e}}{K_\mathrm{i}} = Z^2\sqrt{\frac{m_\mathrm{i}}{m_\mathrm{e}}}.$$

But generally

- $Z \geq 1$ and $m_\mathrm{i} >> m_\mathrm{e}$;
- Therefore, $K_\mathrm{e} >> K_\mathrm{i}$ and *conduction by electrons* is much more important than *conduction by the ions*.
- This is just a mathematical statement that
  - there are *more electrons* and
  - they *move faster* relative to the ions.
- Thus, *electrons are more efficient than ions* at transporting heat.

In summary, the *current produced by conduction* is given approximately by

$$j(x) = K_\mathrm{c}\frac{dT}{dx},$$

where $K_\mathrm{c} \simeq K_\mathrm{e}$ is *dominated by the electronic contribution*.

7.4 Radiative Energy Transport by Photons

Assuming stars to *radiate as blackbodies*,

- the *photons* may be viewed as constituting a *relativistic, bosonic gas* with

$$\langle v \rangle = c \qquad u = aT^4 \qquad C = \frac{du}{dT} = 4aT^3.$$

- Thus, by analogy with earlier equations, *for radiative diffusion* we may write

$$j(x) = -K_{\mathrm{r}} \frac{dT}{dx}$$

- where the *coefficient of radiative diffusion* is

$$K_{\mathrm{r}} \equiv \frac{1}{3} \langle v \rangle \lambda C = \frac{4}{3} c \lambda a T^3.$$

  - All quantities are assumed known except the *mean free path* λ.
  - We must now consider various contributions to the *scattering of photons* that are responsible for their effective mean free path in stellar environments.

7.4.1 Thomson Scattering

At high T and low density, *Thomson electron scattering* (scattering of EM radiation by charged particles) dominates

- The cross section is *independent of frequency and T*,

$$\sigma_{\mathrm{T}} = \frac{8\pi}{3}\left(\frac{e^2}{m_{\mathrm{e}}c^2}\right)^2 = 6.652\times 10^{-25}\,\mathrm{cm}^2,$$

- which is valid if $kT << m_{\mathrm{e}}c^2 \quad \rightarrow \quad T << 6\times 10^9\,\mathrm{K}$.
- The corresponding *mean free path* is

$$\lambda_{\mathrm{T}} = \frac{1}{n_{\mathrm{e}}\sigma_{\mathrm{T}}}.$$

- Inserting this in $K_{\mathrm{r}} = \frac{4}{3}c\lambda aT^3$, we obtain for the *coefficient of radiative diffusion* for Thomson scattering

$$K_{\mathrm{r}} \simeq K_{\mathrm{T}} \equiv \frac{acT^3}{2\pi n_{\mathrm{e}}}\left(\frac{m_{\mathrm{e}}c^2}{e^2}\right)^2.$$

- *Assuming Thomson scattering to dominate*, the ratio of coefficients for radiative and conductive transport is

$$\frac{K_{\mathrm{r}}}{K_{\mathrm{e}}} \simeq \frac{K_{\mathrm{T}}}{K_{\mathrm{e}}} = \sqrt{3}Z\frac{P_{\mathrm{r}}}{P_{\mathrm{e}}}\left(\frac{m_{\mathrm{e}}c^2}{kT}\right)^{5/2},$$

  where the *radiation and ideal gas electron pressures* are

$$P_{\mathrm{r}} = \tfrac{1}{3}aT^4 \qquad P_{\mathrm{e}} = n_{\mathrm{e}}kT,$$

  with n_{e} the *electron number density*.

Example: The equation

$$\frac{K_\mathrm{r}}{K_\mathrm{e}} \simeq \frac{K_\mathrm{T}}{K_\mathrm{e}} = \sqrt{3}Z\frac{P_\mathrm{r}}{P_\mathrm{e}}\left(\frac{m_\mathrm{e}c^2}{kT}\right)^{5/2}$$

yields $K_\mathrm{r}/K_\mathrm{e} \simeq 2\times 10^5$ for the Sun.

- This supports our earlier assertion that *radiative transport dominates over conduction* in normal stars.
- This conclusion is based on the assumption of *pure Thomson scattering*,
- but will not be altered significantly by additional photon absorption processes that we consider below.
- However, it is no longer true if the matter in a star has a degenerate equation of state.

7.4.2 Conduction in Degenerate Matter

Electronic conduction in degenerate matter is altered in several important ways relative to that for an ideal gas:

1. Degeneracy typically increases the electron speed by a factor $(\varepsilon_{\rm F}/kT)^{1/2}$, and
2. decreases the heat capacity by a factor of roughly $kT/\varepsilon_{\rm F}$, where $\varepsilon_{\rm F}$ is the fermi energy.
3. The mean free path λ is increased.
4. This is because the exclusion principle allows an electron to scatter to a state only if that state is not already occupied.

The net effect of these changes is that

- *degenerate matter behaves much like a metal* and
- transport of energy by *conduction* becomes important.

We shall return to the issue of conduction in degenerate matter when we consider the structure of white dwarf stars in later chapters.

7.4.3 Absorption of Photons

In addition to simple Thomson scattering of photons, they may be absorbed.

1. Simultaneous *conservation of energy and momentum* prohibits such absorptions on free electrons.
2. However, they are permissible for electrons in the vicinity of ions.
3. Thus, absorption will generally become more important at *higher densities and lower temperatures.*

The two most important absorptive processes in stars are

1. *bound–free absorption*, where
   - The electron that the photon interacts with is initially bound to an ion and is ejected by the interaction.
   - This process is also called *photo-ionization*.
2. *free–free absorption*, where
   - The electron is unbound before and after the interaction.
   - This process is also called *inverse bremsstrahlung*.

Unlike the case for Thomson scattering, both classes of absorptive processes (free–free and bound–free) imply a *frequency-dependent mean free path.*

- In the frequency range ν to $\nu + d\nu$, the photon *energy density* and *heat capacity* are given by

$$u_\nu d\nu = \frac{8\pi}{c^3}\left(\frac{h\nu^3}{e^{h\nu/kT} - 1}\right) d\nu \qquad C_\nu d\nu = \frac{\partial u_\nu}{\partial T} d\nu.$$

- Let λ_ν be the *mean free path* for photons at frequency ν.
- Then the total *coefficient of radiative transport* is obtained by integration:

$$K_\mathrm{r} = \frac{1}{3}\int_0^\infty \langle v\rangle \lambda_\nu C_\nu \, d\nu = \frac{c}{3}\int_0^\infty \lambda_\nu C_\nu \, d\nu.$$

- Introducing the *Rosseland mean,*

$$\lambda_\mathrm{Ross} \equiv \frac{1}{4aT^3}\int_0^\infty \lambda_\nu C_\nu \, d\nu,$$

  allows the above expression for K_r to be written as

$$K_\mathrm{r} = \frac{4}{3} caT^3 \lambda_\mathrm{Ross}.$$

This is the same form as the previous expressions for frequency-independent photon mean free paths with the replacement $\lambda \rightarrow \lambda_\mathrm{Ross}$.

7.4.4 Stellar Opacities

The total probability of photon interaction is a sum of contributions from *electron and ion scattering*.

- Since $\lambda \sim (n\sigma)^{-1}$, where n is a number density and σ is a cross section, we may write that generally
$$\lambda = \frac{1}{n_{\rm e}\sigma_{\rm e} + n_{\rm i}\sigma_{\rm i}}.$$
- Both the electron number density $n_{\rm e}$ and the ion number density $n_{\rm i}$ are *proportional to the matter density*.
- Thus we may parameterize the preceding equation in the form
$$\lambda = \frac{1}{\rho\kappa},$$
where κ is termed the *opacity*, which has units of *area divided by mass.*

Thus, we may use $\lambda = 1/\rho\kappa$ to rewrite our previous formulas in terms of the opacity κ instead of the mean free path λ.

Mean free path and *opacity* convey the same information, but in the literature opacity is typically the preferred quantitity.

7.4.5 General Discussion of Contributions to Opacity

We may make the following qualitative remarks concerning ionization and the various components of the stellar opacity.

1. *Bound–free absorption* is important at *low temperatures* where atoms are only *partially ionized*.
2. *Free–free absorption* is dominant at *higher temperature* where atoms become *fully ionized*, producing many free electrons with which to interact.
3. *Thomson scattering* contributes a *constant background* that is *independent of temperature.*

- An approximate expression for the frequency-averaged opacity deriving from the free–free and bound–free mechanisms is given by *Kramer's Law:*

  $$\kappa_{\rm ab} = \kappa_0 \rho T^{-3.5},$$

  where "ab" denotes an *absorption-dominated opacity*.

- Approximate formulas for the *free–free and bound–free frequency-averaged absorption opacities* above $T \sim 10^4$ K may be given in the Kramer's form

  $$\kappa_{\rm ff} \simeq 4 \times 10^{22} (X+Y)(1+X) \rho T^{-3.5}\ {\rm cm^2\ g^{-1}}$$
  $$\kappa_{\rm bf} \simeq 4 \times 10^{25} Z (1+X) \rho T^{-3.5}\ {\rm cm^2\ g^{-1}}.$$

- Thomson scattering gives a *constant background opacity*

  $$\kappa_{\rm T} = \frac{1}{\lambda_{\rm T}\rho} = \frac{n_{\rm e}\sigma_{\rm T}}{\rho}.$$

  Introducing for a fully ionized gas

  $$n_{\rm e} \simeq \frac{(1+X)\rho}{2m_{\rm H}} \qquad n_{\rm i} \simeq \frac{(2X+\frac{1}{2}Y)\rho}{2m_{\rm H}},$$

  we may approximate the *Thomson scattering opacity* as

  $$\kappa_{\rm T} \simeq \frac{(1+X)\sigma_{\rm T}}{2m_{\rm H}} = 0.20(1+X)\ {\rm cm^2\ g^{-1}}.$$

Table 7.1: Solar opacities and mean free paths

$R/R_\odot$	T (K)	$\rho(\mathrm{g\,cm^{-3}})$	$\kappa(\mathrm{cm^2\,g^{-1}})$	λ (cm)
0	1.6×10^7	157	1	0.006
0.3	6.8×10^6	12.0	2	0.042
0.6	3.1×10^6	0.50	8	0.25
0.9	6.0×10^5	0.026	100	0.39

Summarizing, the *absorptive and Thomson opacities* may be estimated by

$$\kappa_{\mathrm{ff}} \simeq 4\times10^{22}(X+Y)(1+X)\rho T^{-3.5}\,\mathrm{cm^2\,g^{-1}} \qquad \text{(free–free)},$$

$$\kappa_{\mathrm{bf}} \simeq 4\times10^{25}Z(1+X)\rho T^{-3.5}\,\mathrm{cm^2\,g^{-1}} \qquad \text{(bound–free)},$$

$$\kappa_{\mathrm{T}} \simeq \frac{(1+X)\sigma_{\mathrm{T}}}{2m_{\mathrm{H}}} = 0.20(1+X)\,\mathrm{cm^2\,g^{-1}}. \qquad \text{(Thomson)}$$

- Some realistic opacities calculated for the Sun are given in Table 7.1.
- These opacities indicate that the interior of the Sun is *extremely opaque* to electromagnetic radiation.

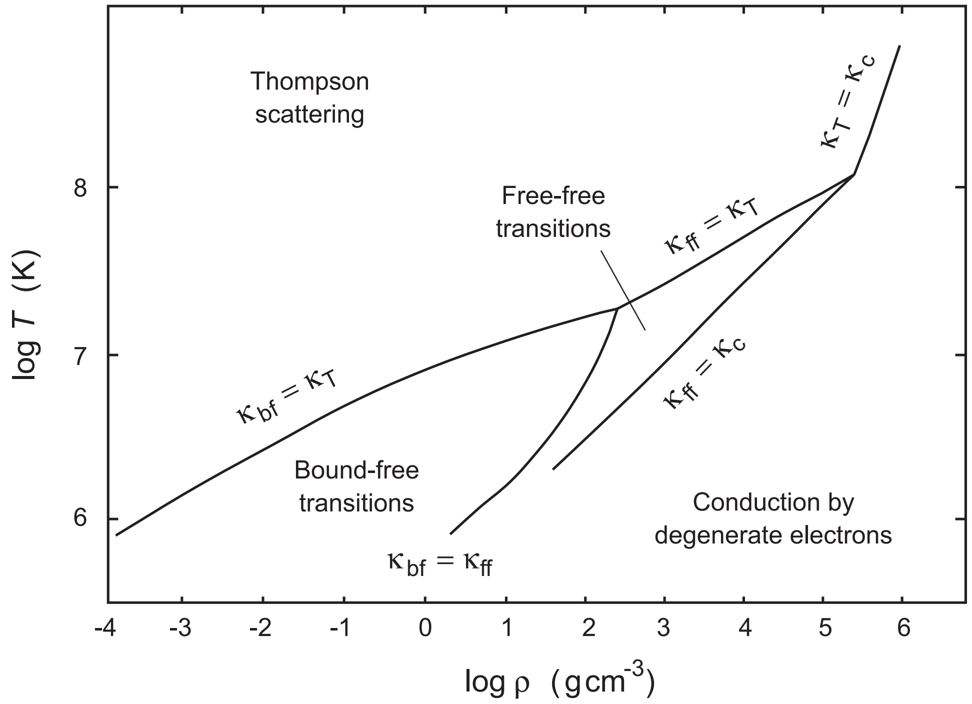


Figure 7.2: Dominant contributions to stellar opacity.

Dominant contributions to the opacity as a function of T and ρ follow from the preceding equations, as illustrated in Fig. 7.2.

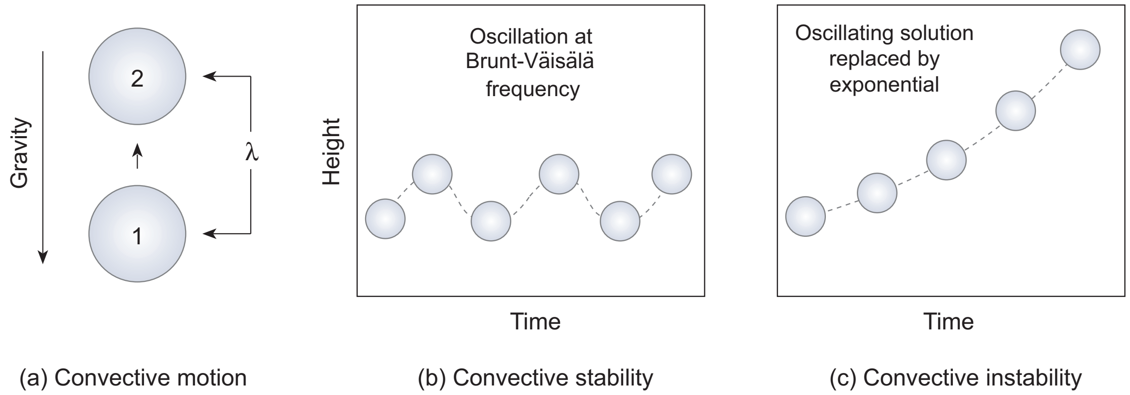
- The boundaries between regions are defined by lines where the corresponding opacities are equal.
- At high temperature and low density Thomson scattering dominates.
- At low temperature and high density electrons are degenerate and matter becomes a good conductor.
- In between, the opacity is dominated by bound–free and free–free transitions.

7.5 Energy Transport by Convection

In some cases the energy to be transported is *too large* to be carried efficiently by radiative transport or conduction.

- Then the system can become *unstable to macroscopic overturn* in a process called *convection*.
- Convection moves entire blobs of material up and down in the gravitational field.
- Thus it can *transport energy very efficiently* when it operates.
- Let us first make a conceptual distinction between two categories of convection.
  1. *Microconvection* applies when the convective blobs are small relative to the region that is unstable.
  2. *Macroconvection* corresponds to convection in which the blobs are a substantial fraction of the size of the convective region.
- This distinction has an important practical implication.
  - Microconvection → spherical symmetry.
  - Macroconvection → spherical symmetry strongly broken (multidimensional hydrodynamics)

- Our initial discussion of convection will be somewhat more general than is normally required for the structure of ordinary stars.
- We do so in order to lay the groundwork for later discussions of events like supernovae in which
  - rapid and complex convective processes may play a significant role, and for which
  - the convection cannot be modeled adequately by theories like the simple *mixing-length theory* described below.


Gravity
2
1
λ
Height
Oscillation at Brunt-Väisälä frequency
Time
Oscillating solution replaced by exponential
Time

(a) Convective motion (b) Convective stability (c) Convective instability

Figure 7.3: (a) Schematic illustration of convective motion. (b) Convectively stable situation: a blob displaced vertically a small amount oscillates around a stable equilibrium with a frequency called the *Brunt–Väisälä frequency*. (c) Convectively unstable situation: a blob displaced vertically continues to rise as time goes on.

7.6 Conditions for Convective Instability

Imagine a blob of matter in a gravitating fluid displaced upward from position 1 to position 2, as illustrated in Fig. 7.3(a).

- If the region is *convectively stable* the displaced blob experiences a restoring force that tends to return it to its original position, as illustrated in Fig. 7.3(b).
- Because of overshooting, the blob executes a stable oscillation around an equilibrium height with a frequency termed the *Brunt–Väisälä frequency.*
- However, if the blob of material at position 2 is *less dense than the surrounding material* it will be driven continuously upward by buoyancy forces.
- The region is then said to be *convectively unstable*, as illustrated in Fig. 7.3(c).

We may choose to impose particular physical conditions on how the blob of matter is moved, and these lead to three separate criteria for convective instability:

- The *Schwarzschild criterion.*
- The *Ledoux criterion.*
- The *double-diffusive criterion.*

We shall now discuss each of these in turn.

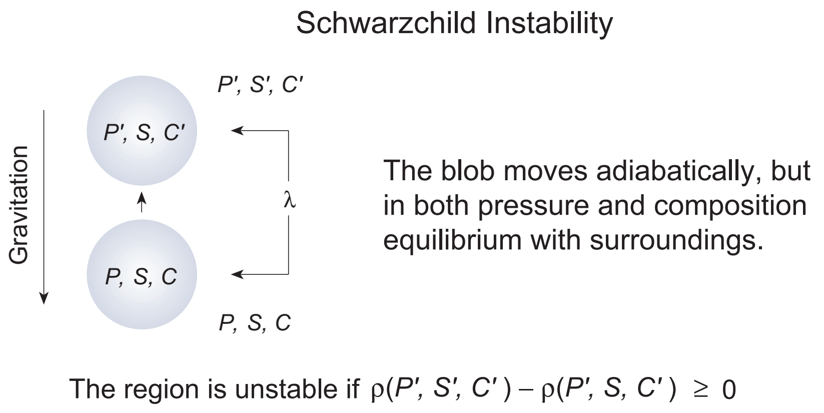

Schwarzchild Instability
P′, S′, C′
P′, S, C′
Gravitation
λ
P, S, C
P, S, C
The blob moves adiabatically, but in both pressure and composition equilibrium with surroundings.
The region is unstable if ρ(P′, S′, C′) – ρ(P′, S, C′) ≥ 0

Figure 7.4: Convective instability by the Schwarzschild criterion.

7.6.1 Schwarzschild Instability

Suppose the blob to move *adiabatically* (constant entropy), but in *pressure and composition equilibrium* with its surroundings.

- Denote the pressure, entropy, and composition of the medium at position 1 by P, S, and C, respectively, and at position 2 by P', S', and C', as illustrated in Fig. 7.4.

- The *condition for convective instability* is that the blob is *less dense than the surrounding medium* at point 2:

$$\underbrace{\rho(P',S',C')}_{\text{Medium }\rho} - \underbrace{\rho(P',S,C')}_{\text{Blob }\rho} \geq 0.$$

  Expanding the density difference in a Taylor series gives

$$\rho(P',S',C') - \rho(P',S,C') = \left.\frac{\partial\rho}{\partial S}\right|_{P,C} \lambda \frac{dS}{dr}.$$

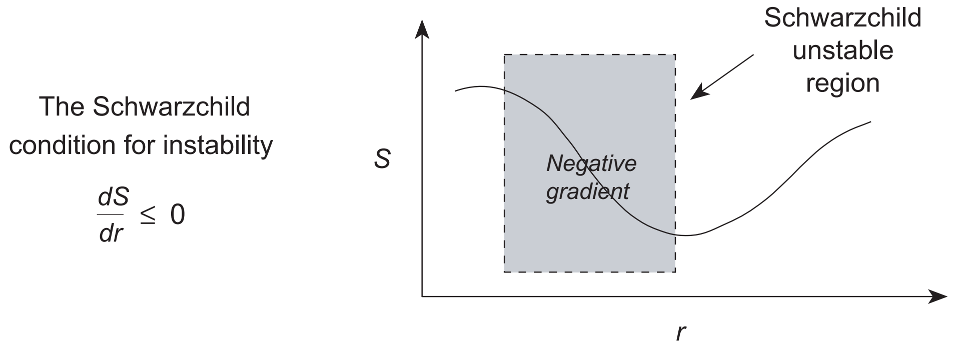

The Schwarzchild condition for instability
dS/dr ≤ 0
S
r
Negative gradient
Schwarzchild unstable region

Figure 7.5: A Schwarzschild-unstable region.

- By using

$$C_P = T\left(\frac{\partial S}{\partial T}\right)_P \quad \longrightarrow \quad \frac{\partial\rho}{\partial S} = \frac{\partial\rho}{\partial T}\frac{\partial T}{\partial S} = \frac{\partial\rho}{\partial T}\frac{T}{C_{\rm P}}$$

  to introduce the heat capacity at constant pressure $C_{\rm p}$, we may exchange the entropy S for the temperature T as a variable and the Schwarzschild condition becomes

$$\underbrace{\rho(P',S',C')}_{\text{Medium }\rho} - \underbrace{\rho(P',S,C')}_{\text{Blob }\rho} \geq 0 \quad \longrightarrow \quad \underbrace{\left(\frac{T}{C_{\rm p}}\frac{\partial\rho}{\partial T}\bigg|_{P,C}\lambda\right)}_{\text{negative}}\frac{dS}{dr} \geq 0.$$

- Typically $\partial\rho/\partial T$ is *negative*, giving the *Schwarzschild condition* for convective instability in the form

$$\frac{dS}{dr} \leq 0 \qquad \text{(Schwarzschild condition)}.$$

- We conclude that a region is unstable against Schwarzschild convection if there is a *negative entropy gradient,* as illustrated schematically in Fig. 7.5.

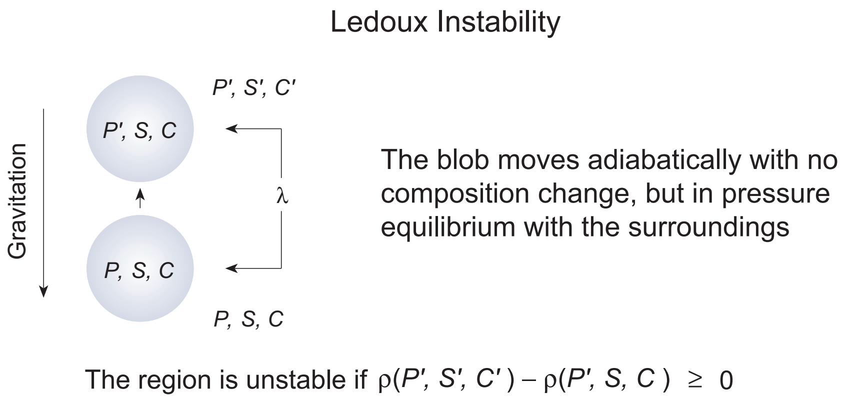

Ledoux Instability
P′, S′, C′
P′, S, C
λ
Gravitation
P, S, C
P, S, C
The blob moves adiabatically with no composition change, but in pressure equilibrium with the surroundings
The region is unstable if ρ(P′, S′, C′) – ρ(P′, S, C) ≥ 0

Figure 7.6: Convective instability according to the Ledoux criterion.

7.6.2 Ledoux Instability

Now suppose that the blob moves *adiabatically with no composition change*, but in *pressure equilibrium,* as in Fig. 7.6.

- The condition for convective instability is now

$$\underbrace{\left(\frac{T}{C_\mathrm{p}}\frac{\partial\rho}{\partial T}\bigg|_{P,C}\lambda\right)\frac{dS}{dr}}_{\text{Schwarzschild}}+\underbrace{\left(\frac{\partial\rho}{\partial C}\bigg|_{P,S}\lambda\right)\frac{dC}{dr}}_{\text{concentration}}\geq 0$$

- The first term is as for Schwarzschild; the second arises because of the assumption of no composition change.

- Usually both partial derivatives are negative and the *Ledoux condition for instability* takes the form

$$\frac{dS}{dr}+b\frac{dC}{dr}\leq 0 \qquad \text{(Ledoux condition)}$$

where b is positive.

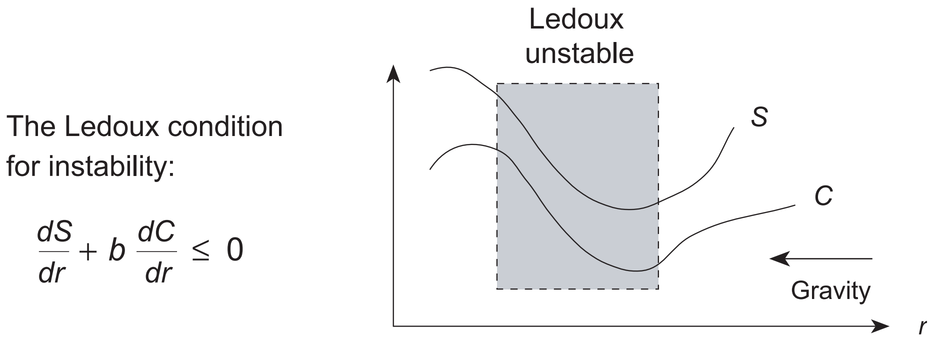
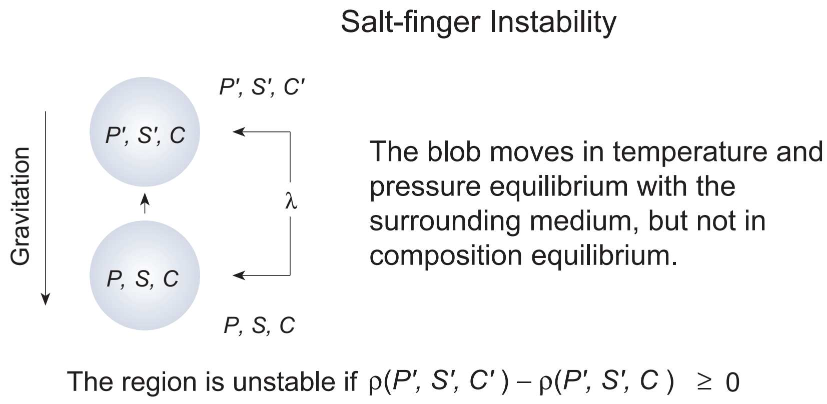


Figure 7.7: A Ledoux-unstable region.

- Therefore, a region is unstable against Ledoux convection if *both the entropy and the concentration variables have negative gradients,* as illustrated schematically in Fig. 7.7.
- If the entropy gradient and concentration gradient have opposite signs, the stability of the region is dependent on the relative sizes of the two terms in

$$\frac{dS}{dr} + b\frac{dC}{dr} \leq 0 \qquad \text{(Ledoux condition)}$$

- For example, a region could be *Schwarzschild-stable but Ledoux-unstable*.



Salt-finger Instability

The blob moves in temperature and pressure equilibrium with the surrounding medium, but not in composition equilibrium.

The region is unstable if ρ(P′, S′, C′) – ρ(P′, S′, C) ≥ 0

Figure 7.8: Convective instability according to the salt-finger criterion.

7.6.3 Salt-Finger (Doubly-Diffusive) Instability

Finally, let us consider a situation where

- the blob is *in temperature and pressure equilibrium* with the surrounding medium, but
- *not in composition equilibrium,* as illustrated in Fig. 7.8.
- The *condition for convective instability* now takes the form,

$$\rho(P',T',C') - \rho(P',T',C) = \left(\left. \frac{\partial \rho}{\partial C} \right|_{P,T} \lambda \right) \frac{dC}{dr} \geq 0.$$

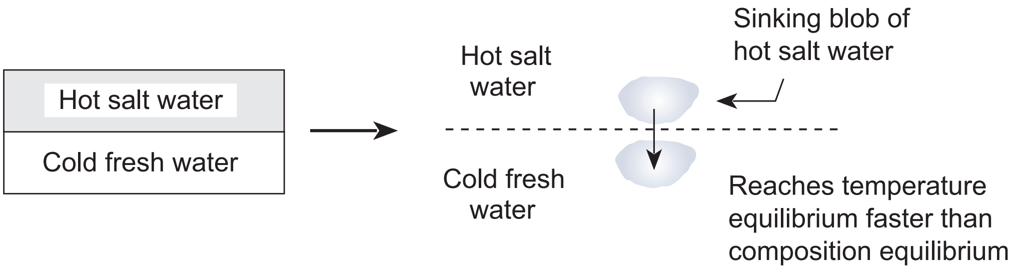

Hot salt water
Cold fresh water
Hot salt water
Cold fresh water
Sinking blob of hot salt water
Reaches temperature equilibrium faster than composition equilibrium

Figure 7.9: Example of salt-finger instability.

We may imagine the following thought experiment in which such an instability could occur.

- Consider a layer of hot salt water that lies over a layer of cold fresh water.
- Now imagine a blob of the hot salt water that begins to sink into the underlying cold fresh water (Fig. 7.9).
- This blob of sinking material will be able to come into heat equilibrium with its surroundings faster than it will be able to come into composition equilibrium.
- This is because transfer of heat by molecular collisions generally is faster than the motion of the sodium and chlorine ions that causes the composition to equilibrate.
- Such a blob may be in approximate temperature equilibrium but remain out of composition equilibrium.
- The heat diffusion will cool the blob of salt water.
- Since salt water is more dense than fresh water at the same temperature, the blob continues to sink.

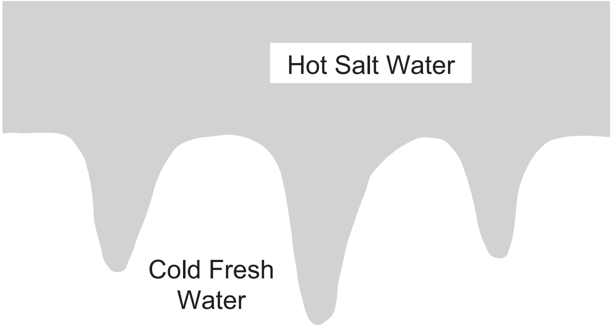


Figure 7.10: Formation of salt fingers.

- As this motion continues, the medium develops “fingers” of salt water reaching down into the fresh water, as illustrated in Fig. 7.10.

This *salt-finger instability* is an example of a class of instabilities that are termed *doubly-diffusive instabilities.* They may occur when

1. Two diffusing substances are present (heat and salt in our example).
2. One of the substances diffuses more rapidly than the other (heat, in our example).
3. The substance diffusing more rapidly has a stabilizing gradient and the slowly diffusing substance has a destabilizing gradient (cold salt water is more dense than cold fresh water).

It is unclear whether such doubly-diffusive instabilities are important in astrophysics. Evidence is not conclusive.

7.7 Critical Temperature Gradient for Convection

- For stars, the most important convective instability is typically that set by the Schwarzschild condition

$$\frac{dS}{dr} \leq 0 \qquad \text{(Schwarzschild condition).}$$

  and driven by entropy gradients.

- The instability criterion for Schwarzschild convection may also be expressed in terms of a critical temperature gradient

$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}},$$

  where the *adiabatic temperature gradient* is defined by

$$\left(\frac{dT}{dr}\right)_{\text{ad}} = \left(1-\frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr} = -\frac{g}{c_{\text{P}}},$$

  and where generally *both derivatives are negative* in the inequality $dT/dr < (dT/dr)_{\text{ad}}$

Thus, a region is convectively unstable if its actual temperature gradient is *steeper than the adiabatic temperature gradient*.

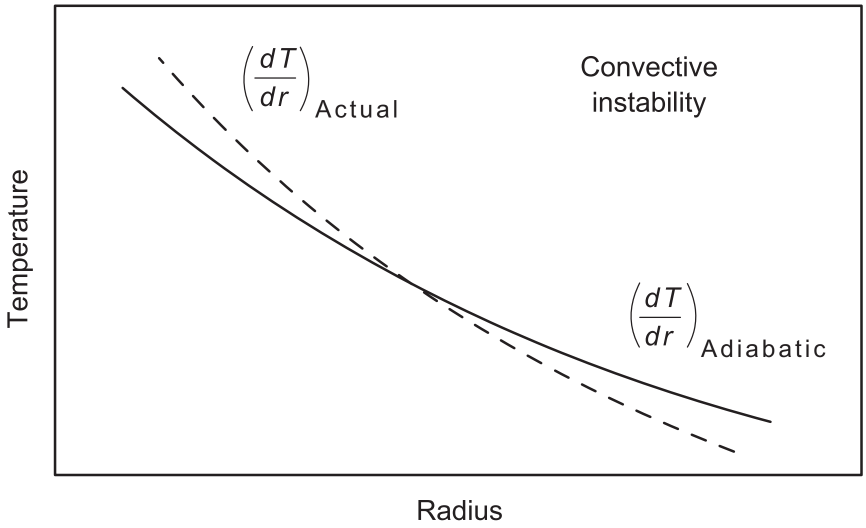


Figure 7.11: Schematic illustration (solid line) of the critical temperature gradient for convection. In this example the actual temperature gradient (dashed line) is steeper than the adiabatic gradient, so the region is convectively unstable.

A schematic illustration of the relationship between the actual temperature gradient and the adiabatic gradient implied by the criterion

$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}},$$

for a Schwarzschild-unstable region is displayed in Fig. 7.11.

- The difference between the actual temperature gradient dT/dr and adiabatic gradient is termed the *superadiabatic gradient* $\delta(dT/dr)$,

$$\delta\left(\frac{dT}{dr}\right) \equiv \frac{dT}{dr} - \left(\frac{dT}{dr}\right)_{\text{ad}}.$$

- Conditions for which this quantity is negative (so that $|dT/dt| > |(dT/dr)_{\text{ad}}|$, since both derivatives are negative) are said to be *superadiabatic.*

- If we divide both sides of

$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}} \qquad \rightarrow \qquad \frac{dT}{dr} < \left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

  by dT/dr (which is negative), we may express the instability condition in the alternative form

$$\frac{d\ln P}{d\ln T} < \frac{\gamma}{\gamma - 1}.$$

- Thus, (Schwarzschild) convective instability requires the temperature to fall off sufficiently fast with height that the actual temperature gradient satisfies

$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}} \qquad \text{or} \qquad \frac{d\ln P}{d\ln T} < \frac{\gamma}{\gamma - 1}.$$

- Equivalently, convective instability implies a *negative superadiabatic gradient*.

$$\frac{dT}{dr} - \left(\frac{dT}{dr}\right)_{\text{ad}} < 0.$$

On the right side of

$$\left(\frac{dT}{dr}\right)_{\text{ad}} = \left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

at given T and P the two most important factors are

1. the *adiabatic index* γ and
2. the *pressure gradient* dP/dr.

Let us now examine in more depth how these factors influence the critical temperature gradient that marks the boundary of convective instability.

7.7.1 Role of the Adiabatic Index in Convection

- For an ideal gas the *adiabatic index* may be expressed as

  $$\gamma = \frac{C_\mathrm{P}}{C_\mathrm{V}} = \frac{1+s/2}{s/2},$$

  where

  - s is the *number of classical degrees of freedom per particle*,
  - each carrying *average thermal energy* $E = \frac{1}{2}kT$.

- Therefore, for a *monatomic gas* with only *three translational degrees of freedom* the adiabatic index is

  $$\gamma = \frac{1+3/2}{3/2} = \frac{5}{3},$$

  and the *condition for convective instability* is that

  $$\frac{d\ln P}{d\ln T} < \frac{\gamma}{\gamma-1} \quad \rightarrow \quad \frac{d\ln P}{d\ln T} < \frac{\frac{5}{3}}{\frac{5}{3}-1} \quad \rightarrow \quad \frac{d\ln P}{d\ln T} < 2.5.$$

- But if the gas has *additional degrees of freedom*,

  - the adiabatic index will *decrease* and
  - for many degrees of freedom it will *approach unity*:

  $$\lim_{s\to\infty} \gamma = \lim_{s\to\infty} \left(\frac{1+s/2}{s/2} \right) = 1.$$

Notice that as $\gamma \to 1$ the factor $(\gamma - 1)/\gamma$ tends to zero and the critical temperature gradient

$$\left(\frac{dT}{dr}\right)_{\text{ad}} = \left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

entering the condition

$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}},$$

becomes *less steep*, since

$$\lim_{\gamma\to 1}\left(\frac{dT}{dr}\right)_{\text{ad}} = \lim_{\gamma\to 1}\left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr} = 0$$

Thus, an *increase in the degrees of freedom* for a gas will generally cause

$$\gamma \longrightarrow 1 \qquad \left(\frac{dT}{dr}\right)_{\text{ad}} \longrightarrow 0$$

thereby *enhancing convective instability.*

Three processes illustrate how an increase in the number of degrees of freedom may occur:

1. Energy may be absorbed by exciting *internal vibrations and rotations of molecules* and emitted by the deexcitation.
2. Energy may be absorbed by the *dissociation of molecules* and emitted in their recombination.
3. Energy may be absorbed by *ionization of atoms or molecules* and emitted in their recombination.

The associated physical process can contribute to convective instability by *increasing the effective number of degrees of freedom* in the gas.

- This *decreases the adiabatic index* toward unity, thereby making the condition

$$\frac{d\ln P}{d\ln T} < \frac{\gamma}{\gamma - 1}$$

easier to fulfill, since

$$\lim_{\gamma \to 1}\left(\frac{\gamma}{\gamma - 1}\right) = \infty.$$

- In later examples we shall see that the physical reason for this decreased convective stability typically is that *these processes permit rising blobs of gas to remain buoyant longer,* thereby enhancing convection.

7.7.2 Role of the Pressure Gradient in Convection

- In hydrostatic equilibrium, the pressure gradient is given by
$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho(r) = -g(r)\rho(r),$$
where $g(r)$ is the local gravitational acceleration.

- Thus, pressure falls off more gradually where $g(r)$ is small and a smaller value of dP/dr make the condition
$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}} \quad \rightarrow \quad \frac{dT}{dr} < \left(1-\frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$
easier to satisfy, thereby favoring convective instability (*Recall: both derivatives are negative*).

We conclude that regions in which the local gravity is weak will be more susceptible to convective instabilities than those with stronger gravity.

Example: In close binary star systems a star may be tidally distorted by its companion. The decreased gravity in tidally distended regions (which occurs because a gas particle there feels a gravitational attraction from one star that is partially canceled by the gravitational attraction of the other star) may initiate convective instability.

7.8 Stellar Temperature Gradients

The condition

$$\frac{dT}{dr} < \left(\frac{dT}{dr}\right)_{\text{ad}} \quad \rightarrow \quad \frac{dT}{dr} < \left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

defines a *critical temperature gradient* for convective instability in terms of the adiabatic temperature gradient

$$\left(\frac{dT}{dr}\right)_{\text{ad}} = \left(1 - \frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

Therefore, we must investigate the *actual* temperature gradients dT/dr of stars in order to assess their stability against convection.

How do we determine the actual temperature gradient in a star?

- Stars will choose the mode of energy transport leading to the *smallest temperature gradient and largest luminosity*.
- The temperature gradients of normal stars that are not convective are determined by the rate of radiative energy transport.

This suggests the following approach for determining the mode of energy transport in nondegenerate stars:

1. Calculate the temperature gradient for radiative transport according to a prescription to be given below.
2. If this gradient is sub-critical, assume no convection and that radiation is the dominant means of energy transport out of the star.
3. If the resulting gradient is critical or supercritical, assume that convection (because it is very efficient in transporting energy) prevails as the means of energy transport as long as the temperature gradient remains critical.

Notice that these considerations could lead to different conclusions in different regions of a star; thus, we expect that

> Stars may be *convective in some regions* and *radiative in others.*

7.8.1 Radiative Gradients

- Let $L(r)$ denote the rate of energy flow through a shell of thickness dr at a radius r, and
- let $\varepsilon(r)$ denote the nuclear power per unit volume generated at radius r.
- Then the power generated in the shell of thickness dr at radius r is given by $4\pi r^2 \varepsilon(r) dr$.
- This is added to the outward power flow from interior shells and
$$\frac{dL}{dr} = 4\pi r^2 \varepsilon(r).$$
- Outside the central power-generating regions for a star we may expect that $L(r)$ approaches a constant equal to the surface luminosity of the star.
- If we assume the energy flow to be radiative,
$$L(r) = 4\pi r^2 j(r),$$
where for radiative transport (earlier result rewritten in terms of the opacity κ)
$$j(r) = -\frac{4acT^3}{3\rho\kappa}\frac{dT}{dr}.$$

Combining the equations,

$$L(r) = 4\pi r^2 j(r) \qquad j(r) = -\frac{4acT^3}{3\rho\kappa}\frac{dT}{dr}$$

we may solve for the *temperature gradient associated with radiative diffusion*,

$$\left(\frac{dT}{dr}\right)_{\text{rad}} \equiv \left(\frac{dT}{dr}\right) = -\frac{3\rho(r)\kappa(r)}{4acT^3(r)}\frac{L(r)}{4\pi r^2}.$$

If this radiative gradient becomes steeper than the critical gradient, the system will become convectively unstable.

Example: Use the preceding results to estimate the temperature gradients in the Sun.

- At $R = 0.3R_\odot$ assume that the luminosity becomes equal to the surface luminosity of $4 \times 10^{33}\,\mathrm{erg\,s^{-1}}$.
- At this radius we have from the Standard Solar Model and opacity tables

$$T \sim 6.8 \times 10^6\,\mathrm{K} \quad \rho \sim 12\,\mathrm{g\,cm^{-3}} \quad \kappa \sim 2\,\mathrm{cm^2\,g^{-1}}.$$

- Then the average solar temperature gradient is

$$\frac{dT}{dr} \simeq -1 \times 10^{-4}\,\mathrm{K\,cm^{-1}},$$

- The average mean free path is

$$\lambda = \frac{1}{\rho\kappa} \simeq 0.04\,\mathrm{cm}$$

- The fractional change in temperature over a distance of one mean free path is of order 10^{-12}.

Thus, we find that

- the solar interior is *extremely opaque,*
- *temperature changes very slowly* over a characteristic diffusion distance,

validating the radiative diffusion assumption.

7.9 Mixing-Length Treatment of Convection

A proper treatment of convection is stars is a difficult subject because it requires the solution of 3-dimensional hydrodynamics for a turbulent, compressible fluid.

- Although modern computing power is making headway on this issue, historically much of our understanding of convection has derived from simple models based on *mixing-length approximations*.
- These models have rather murky theoretical foundation but they appear to work well as phenomenological descriptions of the most important aspects of convection in normal stars.

Part of that success is because of the empirical nature of mixing-length models. Part is because

1. Convection is such an efficient source of energy transport that it often dominates all other modes (so we don't have to think too deeply about partitioning energy transport between radiation and convection).
2. Convection can often operate with convective velocities that are well below sound speed (so that no shock waves are produced)
3. Convection can often operate on a timescale that is well separated from other relevant timescales (such as the hydrodynamic response time) in the star.

However, one should not forget that

- Mixing-length models are basically empirical, with the most essential parameter (the mixing-length) not specified by any fundamental theory.
- As a consequence, they break down in a variety of important cases.

For example, mixing-length models are generally not very appropriate for situations where

1. Radiative transport is competing strongly with convection, such as in the surface of a convective star.
2. Convective transport is supersonic and thus produces shock waves (such mechanisms may operate in the surface of the Sun and contribute to heating of the corona).
3. Convection violates spherical symmetry strongly, as is thought to happen in supernova explosions.
4. The timescales for convection are comparable to other dynamical timescales in the system, as is thought to happen for pulsating variable stars and for supernova explosions.

With this as introduction, and taking due note of the caveats, we now develop a basic mixing-length model of convection.

7.9.1 Pressure Scale Height

Introduce the *pressure scale height* H_p, defined by

$$H_\text{p} \equiv -\frac{dr}{d\ln P} = -P\frac{dr}{dP}.$$

- If H_p is assumed to be constant, the solution of this differential equation is

  $$P = P_0 e^{-r/H_\text{p}}.$$

- The scale height has length dimensions.
- It is the *characteristic vertical scale for variation of the pressure* in an atmosphere since H_p is the vertical distance over which the pressure changes by a factor of e.
- Using the equation for hydrodynamic equilibrium to replace dr/dP and using the ideal gas law, we may express the *scale height for a gravitating ideal gas* as

  $$H_\text{p} = \frac{P}{g\rho} = \frac{kT}{g\mu m_\text{H}},$$

  where g is the local gravitational acceleration, k is Boltzmann's constant, μ is the mean molecular weight, and $m_\text{H} = 1/N_\text{A}$ is the atomic mass unit.

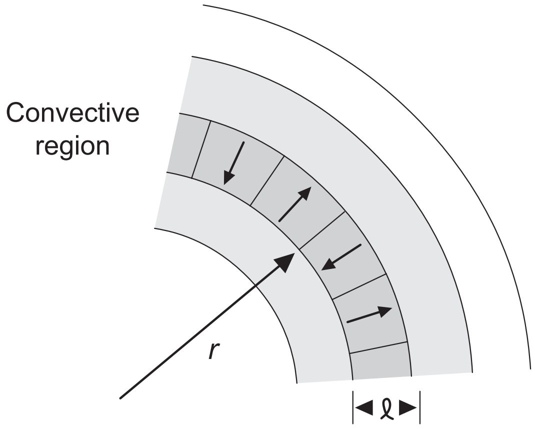


Figure 7.12: Schematic illustration of the mixing-length approximation to convective motion. The mixing length ℓ determines the vertical distance scale over which rising and falling blobs move before merging with the surrounding medium.

7.9.2 The Mixing-Length Philosophy

A mixing-length model assumes that the stellar fluid is composed of blobs that can move vertically in the gravitational field between regions of higher and lower heat content (Fig. 7.12).

- Blobs may move toward the surface because of buoyancy forces, carrying *warmer fluid outward.*
- Meanwhile blobs moving inward because of negative buoyancy carry *cooler fluid inward.*
- There is *no net mass flow* but there is a *outward transport of energy.*
- The characteristic distance over which blobs rise before dissipating is termed the *mixing length,* ℓ.

Mixing-length approaches then analyze the motion of these blobs over a characteristic scale defined by the mixing length with the following general assumptions.

1. Blobs have *dimensions of the same order as the mixing length* ℓ*.*
2. The mixing length ℓ *is much shorter than any other length scale* of physical significance in the star.
3. The *blobs differ only slightly from the surrounding medium* in temperatures, densities, pressure, and composition.
4. The requirement that the internal pressure of blobs remains approximately the same as the surrounding fluid means that the *timescales associated with any processes important in the convection are long* enough that pressure equilibrium is maintained.
5. This implies that the *vertical speeds of the blobs are small* compared with the local speed of sound in the medium.
6. Thus *acoustic and shock phenomena are negligible* for the convection.

Let us now use these assumptions and guidelines to construct a mixing-length model of convection.

7.9.3 A Mixing-Length Model

- Consider a *rising blob in a convective region* described by an *ideal gas equation of state.* In accordance with the preceding discussion, we assume pressure equilibrium.

- By differentiating both sides of the ideal gas law $P = \rho kT/\mu$ we may show that

$$\frac{dP}{P} = \frac{d\rho}{\rho} + \frac{dT}{T}.$$

- But *pressure equilibrium* implies that the left side vanishes and

$$\Delta\rho \simeq -\rho\frac{\Delta T}{T}.$$

- The *buoyancy force per unit volume* f acting on the blob is

$$f = -g\Delta\rho = g\rho\frac{\Delta T}{T}.$$

- But initially ΔT is zero since the blob begins with the same temperature as its surroundings. Thus, the *force averaged over the motion of the blob* may be approximated by

$$\bar{f} \simeq \tfrac{1}{2}g\rho\frac{\Delta T_{\mathrm{f}}}{T},$$

  where ΔT_{f} is the *final temperature difference between the blob and surroundings.*

- The *work* W *done by the buoyancy force* goes into *kinetic energy* $E_{\rm kin}$ *of the blob* (we neglect any viscous forces in this discussion).

- These quantities are given by

$$E_{\rm kin} = \tfrac{1}{2}\rho\bar{v}^2 \qquad W = \bar{f}\ell,$$

  where ℓ is the mixing length and $\bar{v}$ is the *average velocity of the blob.*

- Therefore, *equating the kinetic energy and the work* gives

$$\tfrac{1}{2}\rho\bar{v}^2 = \bar{f}\ell,$$

  which we solve for the average velocity:

$$\bar{v} = \sqrt{\frac{2\bar{f}\ell}{\rho}} = \sqrt{g\ell\frac{\Delta T_{\rm f}}{T}}.$$

The mixing length ℓ is critical but is not specified yet.

- We expect the pressure scale height H_{p} to set the most relevant length scale for the problem:
  - It determines the distance over which there is a substantial change in gas pressure.
  - Our assumption of minimal difference in properties between convective blobs and the surrounding medium would likely not be justified if the mixing length were large measured on this scale.
- Therefore, we *parameterize the mixing length in units of the scale height,*

$$\ell = \alpha H_{\mathrm{p}}.$$

- The quantity α is termed the *mixing-length parameter*. It is assumed adjustable and expected to be of order unity or smaller.
- We shall have to justify, after the fact, whether this choice implies violation of our other assumptions.
- Combining the preceding equations, the *average convective velocity* may then be expressed as

$$\bar{v} = \frac{\alpha k}{\mu m_{\mathrm{H}}} \sqrt{\frac{T}{g} \delta \left(\frac{dT}{dr} \right)}.$$

- The convective flux F_c is given by a product of
  - Temperature difference of blob and surroundings δT,
  - The specific heat at constant pressure c_p (because of our pressure equilibrium assumptions),
  - The density ρ, and
  - The average convective velocity $\bar{v}$:

$$F_\mathrm{c} = \tfrac{1}{2}\rho \bar{v} c_\mathrm{p} \delta T.$$

- Substituting

$$\bar{v} = \frac{\alpha k}{\mu m_\mathrm{H}} \sqrt{\frac{T}{g}\delta\left(\frac{dT}{dr}\right)}.$$

gives

$$F_\mathrm{c} = \frac{\rho c_\mathrm{p} k \alpha}{2\mu m_\mathrm{H}} \sqrt{\frac{T}{g}\delta\left(\frac{dT}{dr}\right)}\,\delta T.$$

But from

$$\ell = \alpha H_\mathrm{p} \qquad H_\mathrm{p} = \frac{kT}{g\mu m_\mathrm{H}},$$

we can write

$$\delta T = \delta\left(\frac{dT}{dr}\right)\ell = \delta\left(\frac{dT}{dr}\right)\alpha H_\mathrm{p} = \frac{\alpha k T}{\mu g m_\mathrm{H}}\delta\left(\frac{dT}{dr}\right),$$

and the convective flux is

$$F_\mathrm{c} = \tfrac{1}{2}\rho \alpha^2 c_\mathrm{p} \left(\frac{k}{\mu m_\mathrm{H}}\right)^2 \left[\frac{T}{g}\delta\left(\frac{dT}{dr}\right)\right]^{3/2}.$$

- The result

  $$F_{\mathrm{c}} = \tfrac{1}{2}\rho\alpha^2 c_{\mathrm{p}} \left(\frac{k}{\mu m_{\mathrm{H}}}\right)^2 \left[\frac{T}{g}\,\delta\left(\frac{dT}{dr}\right)\right]^{3/2}$$

  gives us an approximate expression for the convective flux.

- To use it we must

  1. Choose a value of the phenomenological parameter α.
  2. Determine the difference between the temperature gradient of the blob and its surroundings $\delta(dT/dr)$.

- Solving the preceding equation for this difference gives

  $$\delta\left(\frac{dT}{dr}\right) = \frac{g}{T}\left[2\left(\frac{\mu m_{\mathrm{H}}}{k}\right)^2 \frac{F_{\mathrm{c}}}{\rho c_{\mathrm{p}}\alpha^2}\right]^{2/3}.$$

- In general, if the critical temperature gradient is exceeded the energy transport through a region may involve a combination of radiative and convective transport.

- Therefore, to proceed we must determine the *relative contribution of radiative and convective energy fluxes* in convective regions.

- Assume *all* flux is being carried by convection. Then

$$F_c = \frac{L(r)}{4\pi r^2},$$

where $L(r)$ is the luminosity evaluated at the radius r.

- For this special case of pure convection,

$$\delta\left(\frac{dT}{dr}\right) = \frac{g}{T}\left[\left(\frac{\mu m_H}{k}\right)^2 \frac{L(r)}{2\pi r^2 \rho c_p \alpha^2}\right]^{2/3}.$$

- How superadiabatic does the temperature gradient have to be in order for the preceding equation to be correct (all flux carried by convection)?

- The ratio of the superadiabatic gradient to adiabatic gradient is obtained by dividing the preceding expression by the earlier expression for the adiabatic gradient,

$$\left(\frac{dT}{dr}\right)_{ad} = \left(1-\frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr} = -\frac{g}{c_P} \qquad \text{(Adiabatic gradient).}$$

The result is

$$\frac{\delta(dT/dr)}{|(dT/dr)_{ad}|} = \frac{c_p^{1/3}}{T}\left[\left(\frac{\mu m_H}{k}\right)^2 \frac{L(r)}{2\pi r^2 \rho \alpha^2}\right]^{2/3}$$

Let us now use the preceding equations to make some quantitative estimates for the case of subsurface convection in the Sun.

Example: With the assumption $\alpha = 1$, we find that for the Sun

$$\left(\frac{dT}{dr}\right)_{\rm ad} \simeq 1.6 \times 10^{-4}\,{\rm K\,cm^{-1}},$$

$$\delta\left(\frac{dT}{dr}\right) = 1.8 \times 10^{-10}\,{\rm K\,cm^{-1}},$$

$$\frac{\delta(dT/dr)}{|(dT/dr)_{\rm ad}|} = 1.1 \times 10^{-6},$$

$$\bar{v} = 1 \times 10^4\,{\rm cm\,s^{-1}} = 0.1\,{\rm km\,s^{-1}},$$

This velocity is much less than the local speed of sound, $v_{\rm s} = 2.1 \times 10^7\,{\rm cm\,s^{-1}}$. The corresponding mixing length is

$$\ell = \alpha H_{\rm p} = H_{\rm p} = 4.7 \times 10^4\,{\rm km},$$

and the timescale for the blob to travel the mixing-length distance $\alpha H_{\rm p} = H_{\rm p}$ is

$$t = \frac{H_{\rm p}}{\bar{v}} = 4.7 \times 10^5\,{\rm s} \simeq 5.4\,{\rm days}.$$

Observations and calculations suggest that

- the convection zone is ∼200,000 km thick,
- with characteristic convection cell size $\sim 10^4$ km at the base and $\sim 10^3$ km at the top.

The results of the preceding example support our earlier assertion:

Convection is often such an efficient process that a temperature gradient only slightly steeper than the adiabatic one is sufficient to carry all flux convectively.

- Certainly for the Sun the above analysis suggests that the temperature gradient in convective regions can be *well approximated by the adiabatic gradient*.
- The earlier reservations should be kept in mind, however.

Example: Very near the surface of a star temperature gradients in convective regions may differ substantially from the adiabatic gradient and a slightly superadiabatic gradient may be a very poor approximation for the actual temperature gradient.

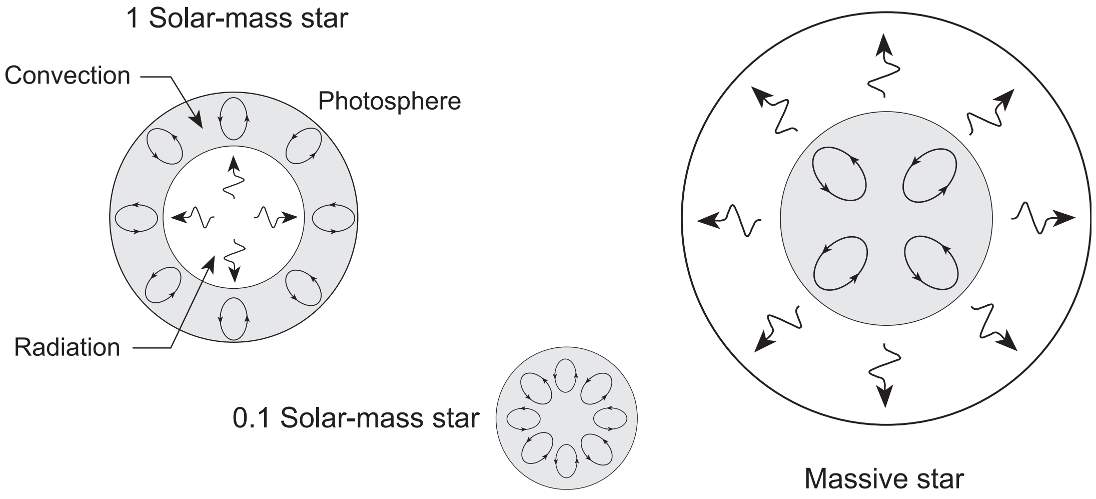


Figure 7.13: Characteristic regions of convection in stars of different mass.

7.10 Examples of Stellar Convective Regions

We expect convection to dominate radiative transport as soon as the critical temperature gradient is reached in a region of a star. Generally, it is believed that (see Figure 7.13)

- The most massive stars are centrally convective and radiative in their outer envelopes.
- Stars of a solar mass or so have subsurface convection zones but the central region is not convective.
- The least-massive stars are entirely convective.

Let's examine two such regions in more detail:

1. *Stellar cores* for massive main sequence stars.
2. *Ionization zones* in the surface layers of stars.

7.10.1 Convection in Stellar Cores

Convection in the cores of stars is favored if the power is generated in a compact central region:

- There is a large energy flow through a region with small gravitational acceleration.
- gravity is weak (little enclosed mass at small radii), so
  - Pressure falls off gradually in this region.
  - Thus rising gas tends to remain buoyant because it need not expand much to pressure-equilibrate.

Let us set

$$\left(\frac{dT}{dr}\right)_{\rm rad} = \left(\frac{dT}{dr}\right)_{\rm ad} \rightarrow -\frac{3\rho(r)\kappa(r)}{4acT^3(r)}\frac{L(r)}{4\pi r^2} = \left(1-\frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

utilize hydrostatic equilibrium to substitute

$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho(r),$$

and rearrange the resulting expression to obtain

$$\frac{L(r)}{m(r)} = \frac{16\pi aGc}{3\kappa}\left(\frac{\gamma-1}{\gamma}\right)\left(\frac{T^4}{P}\right).$$

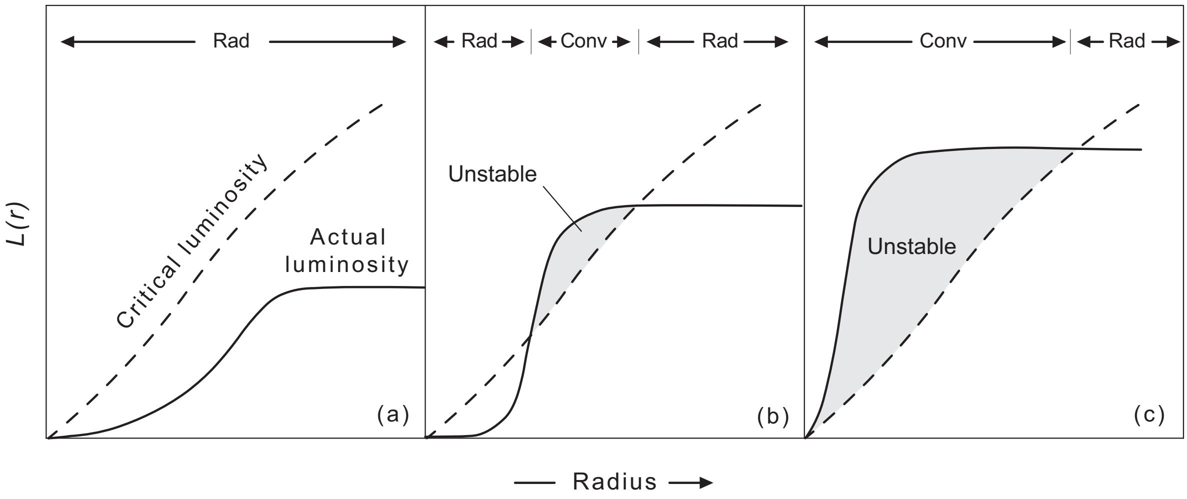


Figure 7.14: Schematic illustration of the competition between radiative and convective energy transport for three qualitatively different situations. The critical luminosity as a function of the radius r is illustrated by the dashed lines and the actual luminosities by the solid lines. For each case, the star is convectively unstable in the shaded regions.

The result

$$\frac{L(r)}{m(r)} = \frac{16\pi aGc}{3\kappa}\left(\frac{\gamma-1}{\gamma}\right)\left(\frac{T^4}{P}\right)$$

defines a critical value of $L(r)/m(r)$ favoring convection over radiative diffusion.

- Generally, we expect convection to develop for any regions of a star in which the luminosity reaches the critical value (which depends on location in the star).
- Some possibilities are indicated schematically in Fig. 7.14.

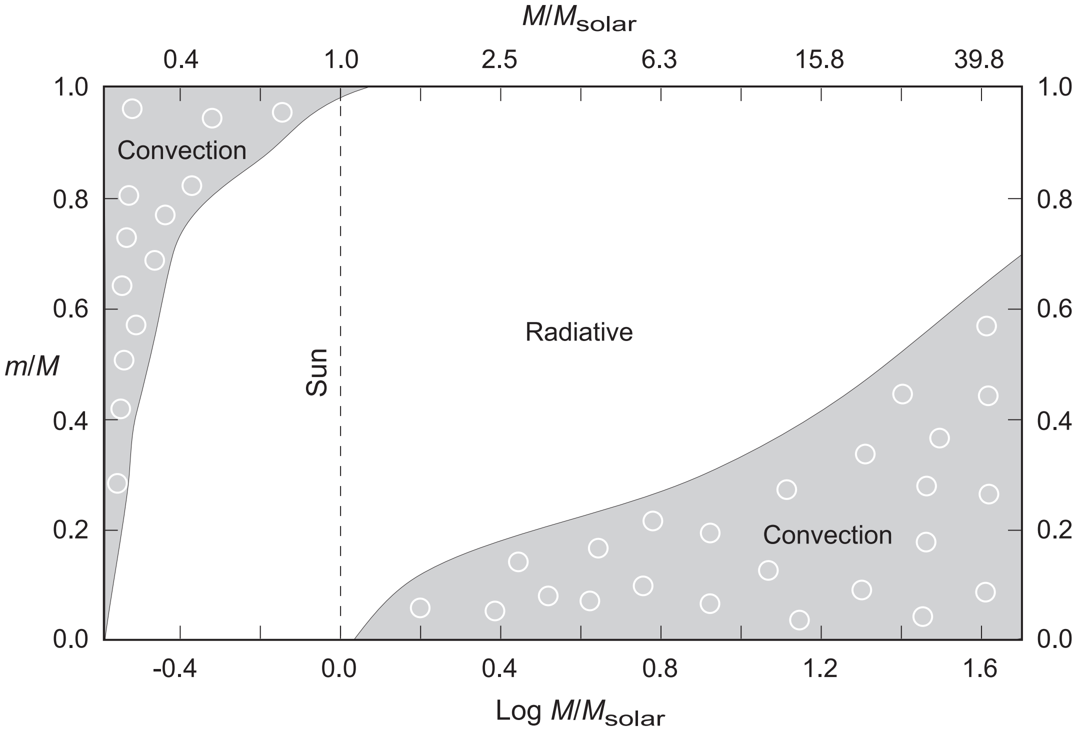


Figure 7.15: Radial extent of convective zones in main sequence stars as a function of stellar mass. The vertical axis is in Lagrangian units of enclosed mass. The position of the Sun is indicated. Figure adapted from Kippenhahn and Wiegert.

A realistic simulation of convective regions as a function of total stellar mass is displayed in Fig. 7.15

Of immediate interest is the suggestion that convective cores of radius r and enclosed mass $m(r)$ can develop in stars if the critical value of $L(r)/m(r)$ defined by

$$\frac{L(r)}{m(r)} = \frac{16\pi aGc}{3\kappa}\left(\frac{\gamma-1}{\gamma}\right)\left(\frac{T^4}{P}\right)$$

is exceeded inside the radius r.

- Such convective cores tend to develop in main sequence stars that are more massive than the Sun.
- In these stars the CNO cycle dominates the energy production mechanism and the strong temperature dependence (power varying as $\sim T^{17}$) confines power production to a small central region.
- For less massive stars like the Sun where the PP chain is the dominant energy production mechanism, the temperature dependence is much weaker (power varying as $\sim T^4$).
- In these stars the energy production is spread over a larger central region and core convection is less likely.

Example: In the case of the Sun inside its 10% radius, data taken from the Standard Solar Model and standard opacity tables allow us to estimate a critical value (Exercise)

$$\frac{L(r)}{m(r)} \simeq 21\,\mathrm{erg\,g^{-1}\,s^{-1}}.$$

Calculations within the *Standard Solar Model* indicate that the Sun's core produces about $11.3\,\mathrm{erg\,g^{-1}\,s^{-1}}$ within its 10% radius.

- This is near but still below the critical value.
- Therefore, we conclude that *the core of the Sun is not convective.*
- However, the centers of main sequence stars more massive than the Sun are likely to be convective.

7.10.2 Surface Ionization Zones

Convection is favored in surface layers, where constant ionization and recombination is taking place, for two reasons

- Opacity is large, making $(dT/dr)_{\text{rad}}$ steep:

$$\left(\frac{dT}{dr}\right)_{\text{rad}} = -\frac{3\rho(r)\kappa(r)}{4acT^3(r)}\frac{L(r)}{4\pi r^2}.$$

- The critical temperature gradient for convection

$$\left(\frac{dT}{dr}\right)_{\text{ad}} = \left(1-\frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr},$$

  is not steep because there are many degrees of freedom s associated with the ionization–recombination and

$$\gamma = \frac{1+s/2}{s/2}$$

  is decreased toward unity.

- More physically, *electron recombination supplies energy to expand the rising gas packets.*

- Thus, packets don't cool much and remain buoyant.

In the Sun, the convective layer from about $0.7R_\odot$ to $0.9R_\odot$ is associated with such ionization zones. This subsurface convection gives rise to the *granules* observed on the solar surface.

7.11 Energy Transport by Neutrino Emission

The cores of massive stars late in their lives become extremely dense and hot.

- These conditions make it difficult to transport their large energy production that is concentrated in a very small region by radiative or even convective processes.
- On the other hand, this very dense, very hot environment is favorable to the production of neutrinos, which by virtue of their weak interactions with matter can leave the core relatively unimpeded.

As a result, neutrino emission generally is the dominant mechanism for cooling stellar cores that proceed beyond carbon burning.

In normal stars, as we have noted above,

- The neutrino emission is not related to the temperature gradient.
- Thus the energy outflow from neutrino cooling is directly proportional to the rate at which the neutrinos are produced in the core of the star.
- As Clayton has stated succinctly,

> *As far as stellar structure and evolution are concerned, neutrinos mostly play the role of a local refrigerator.*

As a rough estimate, the interaction of electron neutrinos with matter has a cross section approximately given by

$$\sigma_\nu \simeq 10^{-44} \left(\frac{E_\nu}{m_e c^2} \right)^2 \text{cm}^2,$$

where E_ν is the neutrino energy and $m_e c^2 = 511\,\text{keV}$ is the electron rest mass energy.

Example: For most processes of importance in stars $E_\nu/m_e c^2$ differs by less than a factor of 10 from unity.

- Thus as a very crude approximation

$$\sigma_\nu \simeq 10^{-44} \left(\frac{E_\nu}{m_e c^2}\right)^2 \mathrm{cm}^2 \quad \rightarrow \quad \sigma_\nu \simeq 10^{-44}\,\mathrm{cm}^2.$$

- Assuming the average density of matter in a representative star to be

$$\rho \simeq 1\,\mathrm{g\,cm}^{-3} \quad \rightarrow \quad n \simeq 10^{24}\,\mathrm{cm}^{-3}$$

  the mean free path for an electron neutrino in average stellar matter is

$$\begin{aligned}\lambda &= \frac{1}{\sigma_\nu n} \\ &\simeq \frac{1}{(10^{-44}\,\mathrm{cm}^2) \times (10^{24}\,\mathrm{cm}^{-3})} \\ &\simeq 10^{20}\,\mathrm{cm} \\ &= 1.4 \times 10^9\, R_\odot.\end{aligned}$$

- Obviously there is little chance that the neutrino scatters from the matter on its way out of a normal star.

7.11.1 Neutrino Production Mechanisms

There are several neutrino production mechanisms that influence the evolution of massive stars. I won't cover them here but they are summarized in the book Chapter. We shall concentrate on the three most important:

- *Pair annihilation,*
- *Photoneutrinos,*
- *Plasma neutrinos.*

Pair production: Neutrino–antineutrino pairs can be produced by the reaction

$$e^- + e^+ \rightarrow \nu_\mathrm{e} + \bar{\nu}_\mathrm{e}.$$

The positrons can be produced in abundance by

$$\gamma + \gamma \rightarrow e^+ + e^-,$$

if $kT \sim m_\mathrm{e}c^2$, implying $T \sim 6 \times 10^9$ K or greater.

Photoneutrinos: When the energy is too low to produce significant numbers of neutrinos by pair production, neutrinos may still be produced by the reaction

$$e^- + \gamma \rightarrow e^- + \nu_\mathrm{e} + \bar{\nu}_\mathrm{e}.$$

Such neutrinos are called *photoneutrinos.* Photoneutrino production generally increases with temperature at all densities.

Plasma neutrinos: At large stellar densities a photon can interact with the plasma to form a collective excitation called a *plasmon*.

- Direct free-space decay of a photon to a neutrino–antineutrino pair is *forbidden by energy and momentum conservation.*

- A *plasmon* is a kind of *"heavy photon"* that acquires an *effective mass* ω_0 through interaction with the medium. The plasmon dispersion relation (for nondegenerate gas) is

  $$\omega^2 = k^2c^2 + \omega_0^2 \qquad \omega_0^2 = \frac{4\pi n_e e^2}{m_e},$$

  where ω_0 is the characteristic *plasma frequency* and k the wavenumber and ω the frequency for the electromagnetic wave.

- A plasmon ("heavy photon") $\gamma_{\rm pl}$ can decay directly to neutrino–antineutrino pairs:

  $$\gamma_{\rm pl} \rightarrow \nu_e + \bar{\nu}_e.$$

- Plasmon neutrino production is important when $\hbar\omega_p \geq kT$. Therefore, plasma neutrino emission is enhanced by *high temperature and high density.*

Table 7.2: Photon and neutrino luminosities for a 20 solar mass star

Fuel	$\rho_c(\,\mathrm{g\,cm^{-3}})^\dagger$	$T_c(10^9\,\mathrm{K})^\dagger$	Duration (y)	$L_\gamma(\,\mathrm{erg\,s^{-1}})$	$L_\nu(\,\mathrm{erg\,s^{-1}})$
Hydrogen	5.6	0.040	1.0×10^7	2.7×10^{38}	—
Helium	9.4×10^2	0.19	9.5×10^5	5.3×10^{38}	$<1.0\times10^{36}$
Carbon	2.7×10^5	0.81	300	4.3×10^{38}	7.4×10^{39}
Neon	4.0×10^6	1.7	0.38	4.4×10^{38}	1.2×10^{43}
Oxygen	6.0×10^6	2.1	0.50	4.4×10^{38}	7.4×10^{43}
Silicon	4.9×10^7	3.7	0.0055	4.4×10^{38}	3.1×10^{45}

†The columns ρ_c and T_c give the critical density and critical temperature to ignite a fuel, respectively.

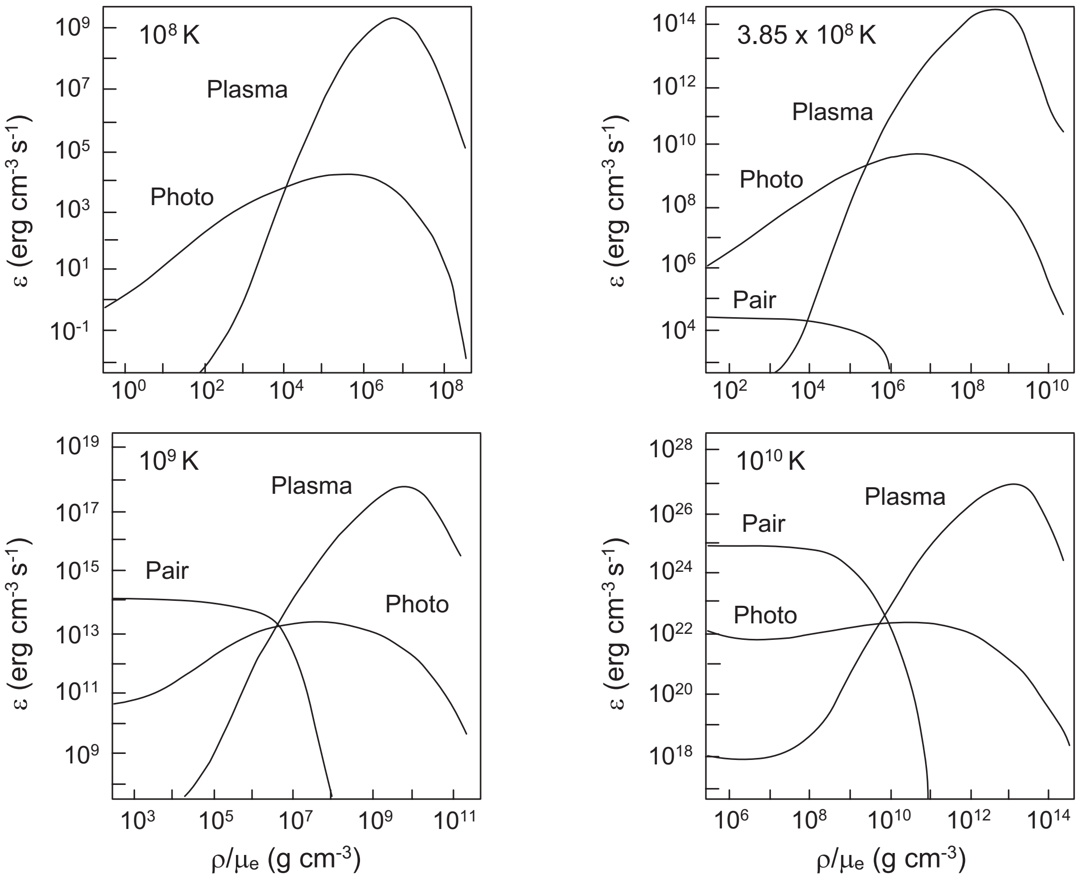


Figure 7.16: Neutrino energy emission rates at four different temperatures.

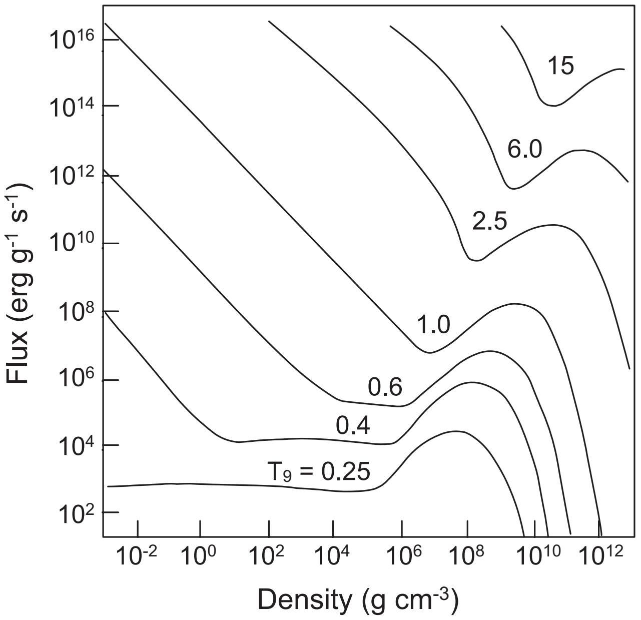


Figure 7.17: Total energy production rates by neutrino processes.

Example: Typical conditions for silicon burning are

$$T = 3-5\times 10^9\,\mathrm{K} \qquad \rho = 10^5 - 10^7\,\mathrm{g\,cm^{-3}}.$$

From Fig. 7.17 the total neutrino energy production rates under these conditions are

$$\varepsilon_\nu \simeq 10^{12} - 10^{15}\,\mathrm{erg\,g^{-1}\,s^{-1}},$$

which is comparable to the Si-burning energy generation rate.

The primary source of core cooling in these late stages of stellar evolution is from neutrino emission.

7.11.2 Coherent Neutrino Scattering

The *Standard Electroweak Theory* of elementary particle physics predicts that

- *neutral weak currents* (those mediated by the Z^0 gauge boson) can scatter coherently off the A nucleons of a composite nucleus rather than off individual nucleons.
- The usual neutrino–nucleon scattering cross section is

  $$\sigma_{\text{nucleon}} \propto E_\nu^2,$$

  where E_ν is the neutrino energy.
- But the *coherent cross section* on a nucleus of nucleon number A is

  $$\sigma_{\text{coherent}} \propto A^2 E_\nu^2.$$

- For massive stars, Si burning produces iron-group nuclei and the coherent cross section is *enhanced by a factor*

  $$\frac{\sigma_{\text{coherent}}}{\sigma_{\text{nucleon}}} \sim A^2 \sim (56)^2 \sim 3000$$

  relative to normal nucleonic weak interactions (taking a mass of 56 amu as representative of iron-group isotopes).
- This enhancement, coupled with the large increase in the normal weak interactions strength because of the enormous temperature and density of the core, implies very large neutrino interactions.

- Because of the large mass difference between neutrinos and heavy nuclei, coherent scattering transfers momentum but little energy, so it is nearly elastic.
- We shall see in later chapters that coherent elastic scattering of neutrinos from composite nuclei through the neutral weak current can have a large influence on the core collapse in a core collapse supernova.

Chapter 8

Summary of Stellar Equations

Two equations governing *hydrostatic equilibrium*,

$$\frac{dm}{dr} = 4\pi r^2 \rho(r) \qquad \text{Mass conservation}$$

$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho \qquad \text{Hydrostatic equilibrium,}$$

three equations for *luminosity and temperature gradients*,

$$\frac{dL}{dr} = 4\pi r^2 \varepsilon(r) \qquad \text{Luminosity}$$

$$\frac{dT}{dr} = -\frac{3\rho(r)\kappa(r)}{4acT^3(r)}\frac{L(r)}{4\pi r^2} \qquad \text{Radiative } T \text{ gradient}$$

$$\frac{dT}{dr} = \left(\frac{\gamma-1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr} \qquad \text{Convective } T \text{ gradient,}$$

equations governing *nucleosynthesis*,

$$\frac{dn}{dt} = -\frac{1}{\tau}n \qquad \text{Nucleosynthesis,}$$

and an *equation of state*,

$$P = P(T, \rho, X_i, \ldots) \qquad \text{Equation of state.}$$

The two temperature-gradient equations are to be employed as follows:

- The radiative gradient

$$\left(\frac{dT}{dr}\right)_{\rm rad} = -\frac{3\rho(r)\kappa(r)}{4acT^3(r)}\frac{L(r)}{4\pi r^2}$$

  should be used unless the condition

$$\left(\frac{dT}{dr}\right)_{\rm rad} < \left(1-\frac{1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

  for convective instability is satisfied, in which case the adiabatic gradient should be used:

$$\left(\frac{dT}{dr}\right)_{\rm ad} = \left(\frac{\gamma-1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr}$$

- Some equations in this set, like the last two,

$$\frac{dn}{dt} = -\frac{1}{\tau}n \qquad \text{Nucleosynthesis,}$$

$$P = P(T,\rho,X_i,\ldots) \qquad \text{Equation of state,}$$

  are to be understood schematically.

  - Nucleosynthesis will in general involve a complex coupled network of differential equations
  - The equation of state will depend on the physics of the problem and may take a variety of forms.

These equations represent a considerably simplified description of a star.

- Even in this simplified form their solution for realistic cases presents formidable numerical problems.
- Relatively specialized techniques must be used for some aspects of the solution
    1. because of the boundary conditions required for a star, and
    2. because these equations couple processes having characteristic time and length scales that may differ by many orders of magnitude.

Let us now consider solution of the stellar equations. We shall address three aspects

- Using timescale analysis to avoid solving the equations directly.
- Obtaining solutions by approximating the stellar equations.
- Full numerical solution of the stellar equations.

Table 8.1: Some important stellar timescales

Timescale	Characteristic value	Value for Sun
Dynamical	$\tau_{\rm dyn} \sim \sqrt{\dfrac{R^3}{2GM}}$	55 min
Thermal adjustment	$\tau_{\rm therm} \sim \dfrac{GM^2}{RL}$	3×10^7 yr
Nuclear burning	$\tau_{\rm nuc} \sim \varepsilon \dfrac{Mc^2}{L}$	10^{10} yr

8.1 Summary of Important Stellar Timescales

A timescale τ_s characteristic of some important physical process represented by a quantity s may be defined as $\tau_s = s/\dot{s}$.

This is just a generalization of the standard example from introductory physics of estimating a time to travel some distance x as $t = x/\dot{x} = x/v$, where v is the average velocity.

At several points in the preceding discussion, three important timescales have been discussed. These are summarized in Table 8.1 and discussed further below.

1. *Dynamical timescale:* A dynamical timescale is defined by a characteristic time to restore hydrostatic equilibrium:

   $$\tau_{\mathrm{dyn}} = \frac{R}{v_{\mathrm{esc}}} = \sqrt{\frac{R^3}{2GM}} \sim \sqrt{\frac{1}{G\bar{\rho}}}$$

   where

   $$v_{\mathrm{esc}} = (2GM/R)^{1/2} \qquad \bar{\rho} = 3M/4\pi R^3$$

   were used. For the Sun $\tau_{\mathrm{dyn}} \sim 55$ minutes.

2. *Thermal adjustment timescale:* The thermal adjustment (or Kelvin–Helmholtz) timescale is associated with time for a star to shed thermal energy, so

   $$\tau_{\mathrm{therm}} = \frac{U}{L} = \frac{GM^2}{LR},$$

   where U is the internal energy and L the luminosity, and $U \sim GM^2/R$ by the virial theorem. The Sun has a thermal adjustment timescale of about 3×10^7 yr.

3. *Nuclear burning timescale:* The time to burn the star's nuclear fuel may be approximated by

   $$\tau_{\mathrm{nuc}} = \frac{\varepsilon M_0 c^2}{L},$$

   where $\varepsilon \sim 0.007$ is the efficiency for conversion of mass into energy in hydrogen fusion, M_0 is the mass of hydrogen available to burn in the star. For the Sun this gives $\tau_{\mathrm{nuc}} \sim 10^{10}$ yr.

8.2 An Approximate Solution: The Lane–Emden Equations

The equations of hydrostatic equilibrium may be combined to give the differential equation

$$\left.\begin{aligned} dm &= 4\pi r^2 \rho(r) dr \\ \frac{dP}{dr} &= -\frac{Gm(r)}{r^2}\rho \end{aligned}\right\} \rightarrow \frac{1}{r^2}\frac{d}{dr}\left(\frac{r^2}{\rho}\frac{dP}{dr}\right) = -4\pi G\rho .$$

We then approximate the equation of state in polytropic form,

$$P = K\rho^{\gamma} = K\rho^{1+1/n} \qquad \gamma \equiv 1 + \frac{1}{n}.$$

Introducing dimensionless variables ξ and θ through

$$\rho = \rho_{\mathrm{c}}\theta^n \qquad r = a\xi \qquad a = \sqrt{\frac{(n+1)K\rho_{\mathrm{c}}^{(1-n)/n}}{4\pi G}},$$

where $\rho_{\mathrm{c}} \equiv \rho(r=0)$ is the central density, the differential equation embodying hydrostatic equilibrium for a polytropic equation of state may be expressed in terms of the new *dependent variable* $\theta(\xi)$ and *independent variable* ξ as,

$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2\frac{d\theta}{d\xi}\right) = -\theta^n .$$

In terms of these new variables the boundary conditions are

$$\theta(0) = 1 \qquad \theta'(0) \equiv \left.\frac{d\theta}{d\xi}\right|_{\xi=0} = 0,$$

- The first follows from the requirement that the correct central density $\rho_\mathrm{c} = \rho(0)$ be reproduced.
- The second follows from requiring that the pressure gradient dP/dr vanish at the origin (necessary condition for hydrostatic equilibrium).

Then we may integrate

$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2\frac{d\theta}{d\xi}\right) = -\theta^n.$$

outward from the origin ($\xi = 0$) until the point $\xi = \xi_1$ where θ first vanishes, to define the surface of the star, since at this point $\rho = P = 0$ because

$$\rho = \rho_\mathrm{c}\theta^n \qquad P = K\rho^\gamma.$$

Solutions having this property generally exist for $n < 5$.

Table 8.2: Lane–Emden constants

n	γ	ξ_1	$\xi_1^2\lvert\theta'(\xi_1)\rvert$
0	∞	2.4494	4.8988
0.5	3	2.7528	3.7871
1.0	2	3.14159	3.14159
1.5	5/3	3.65375	2.71406
2.0	3/2	4.35287	2.41105
2.5	1.4	5.35528	2.18720
3.0	4/3	6.89685	2.01824
4.0	5/4	14.97155	1.79723
4.5	1.22	31.83646	1.73780
5.0	1.2	∞	1.73205

- The equation

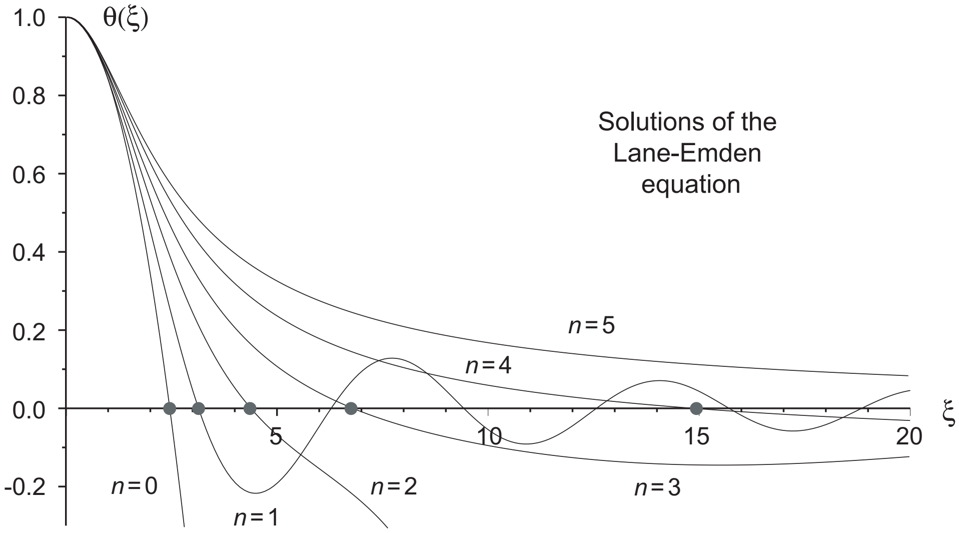
$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2\frac{d\theta}{d\xi}\right) = -\theta^n.$$

  has analytical solutions for the special cases $n = 0, 1,$ and 5, but

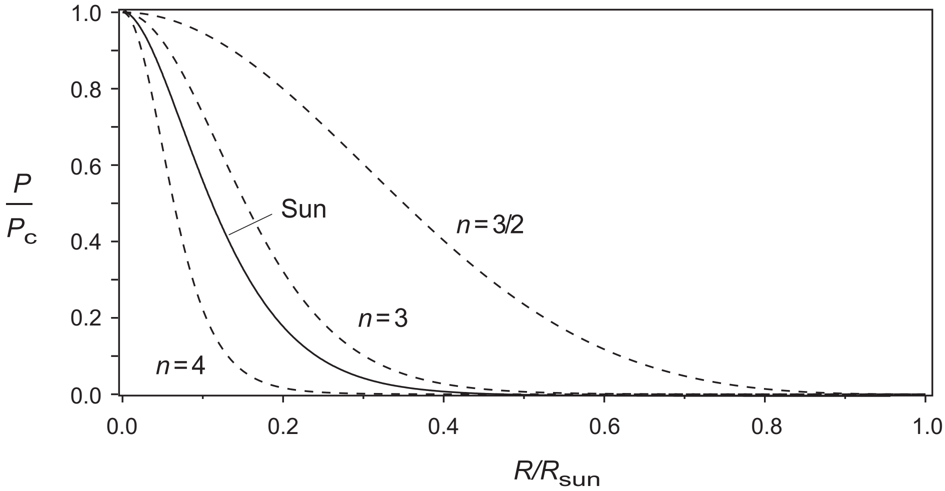
- In the physically most interesting cases the equations must be integrated numerically to define the Lane–Emden constants ξ_1 and $\xi_1^2\lvert\theta'(\xi_1)\rvert$ for given n.

These are tabulated for various values of n and γ in Table 8.2.

Corresponding solutions are plotted in the following figure



and pressure profiles computed for polytropic equations of state with several values of n are shown in the following figure.



The $n = 3$ polytrope approximates relatively well the actual pressure profile of the Sun (Standard Solar Model).

The transformation equations

$$\rho = \rho_c \theta^n \qquad r = a\xi \qquad a = \sqrt{\frac{(n+1)K\rho_c^{(1-n)/n}}{4\pi G}},$$

may then be used to express quantities of physical interest in terms of these constants for definite values of the polytropic index n. For example, the radius R is

$$R = a\xi_1 = \sqrt{\frac{(n+1)K}{4\pi G}}\rho_c^{(1-n)/2n}\xi_1,$$

and the mass M is given by (Exercise)

$$\begin{aligned} M &\equiv 4\pi a^3 \rho_c \left[-\xi^2 \frac{d\theta}{d\xi}\right]_{\xi=\xi_1} \\ &= 4\pi \left[\frac{(n+1)K}{4\pi G}\right]^{3/2} \rho_c^{(3-n)/2n}\xi_1^2|\theta'(\xi_1)|, \end{aligned}$$

Eliminating ρ_c between these two equations gives a general relationship between the mass and the radius,

$$M = 4\pi R^{(3-n)/(1-n)}\left(\frac{(n+1)K}{4\pi G}\right)^{n/(n-1)} \xi_1^{(3-n)/(n-1)}\xi_1^2|\theta'(\xi_1)|.$$

for a solution with polytropic index n.

8.2.1 Limitations of the Lane–Emden Approximation

The Lane–Emden equation has elegant solutions with a direct physical interpretation, but it has serious limitations:

- It reflects only the property of *hydrostatic equilibrium,* and then only for a *polytropic equation of state.*
- Thus it describes only the *mechanical part of stellar structure*.
- It has nothing to say about *temperature gradients* and *energy transport,* and their coupling to the full problem.

There are two general situations where a polytropic equation of state may be reasonable.

- The realistic equation of state depends on T as well as ρ, but additional physical constraints between T and P lead to a polytropic equation of state.

> *Example:* The adiabatic constraint applied to an ideal gas leads to a polytropic equation of state $PV^{\Gamma_1} \propto P\rho^{-\Gamma_1} = \text{constant}$. Then the temperature is effectively fixed by a constraint $T = T(P)$ and not by coupling to the full set of equations.

- The realistic equation of state actually is approximately polytropic. Often true in very dense matter such as white dwarfs and neutron stars.

8.3 Numerical Solution of the Stellar Equations

The stellar structure and evolution problem has some specific features that complicate obtaining numerical solutions. These issues fall primarily into two categories:

- Boundary conditions.

  Some boundary conditions must be imposed at the center and some at the surface. This requires specialized techniques to ensure compatibility of the solutions.

- Extreme space and time scale differences.

  Example: Equations governing isotopic composition and energy release for the PP chains involve timescales that can differ by 10–20 orders of magnitude. They can be solved only with custom numerical methods.

Numerical solution of the full set of equations describing stellar structure and stellar evolution is a specialized topic that would take us too far afield for the present discussion.

Part II

Stellar Evolution

Chapter 9

The Formation of Stars

Substantial direct and indirect information indicates that *stars are born in nebulae.*

- Basics are well understood, many details are not.
- We shall have to gloss over various sticky points with assumptions that will be justified by the observation that *stars exist and, therefore, something like our assumption must be correct.*
- Much of this gloss is associated with the general observation that clouds that collapse to form stars have
  - too much kinetic energy and
  - too much angular momentum

  to produce directly the stars that we see.

Since nature makes stars in abundance, this indicates that there exist mechanisms for nascent stars to shed these excess quantities. It is the details of how this happens that we shall circumvent with appeals to observations.

9.1 O and B Associations and T-Tauri Stars

- Observation of many hot O and B spectral class stars in and near nebulae is a rather strong indicator that stars are being born there.
- These stars are so luminous that they must consume their nuclear fuel at a prodigious rate.
- Their time on the main sequence is probably only a million years or so, therefore they cannot be far from their place of birth.
- We also see, usually in association with stellar O and B complexes in dust clouds, *T-Tauri variables*.
- These are red irregular variables (spectral class F–M), with a number of unusual characteristics. They exhibit emission lines of hydrogen, Ca^+, and some other metals.

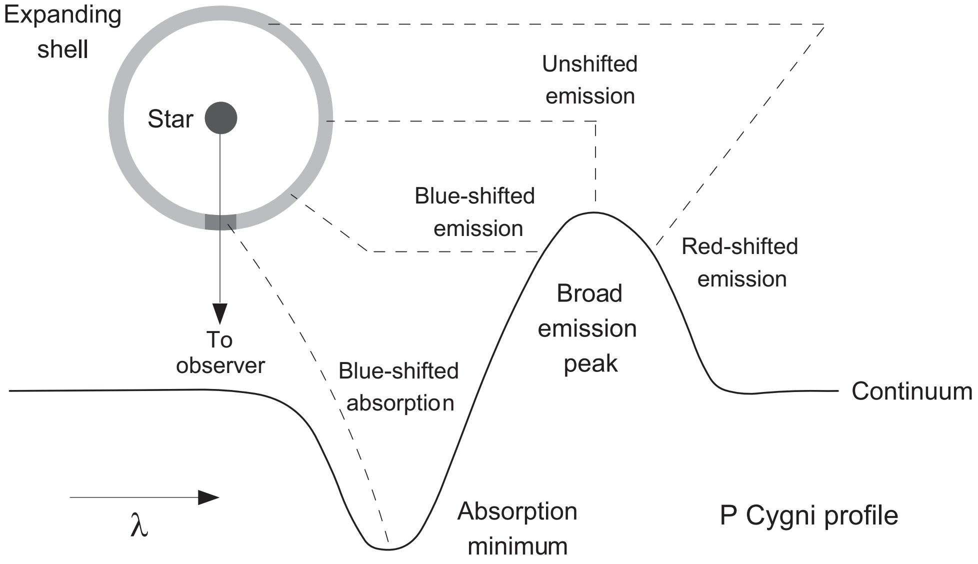


Figure 9.1: Origin of P Cygni profiles in Doppler shifts associated with expanding gas shells.

- The spectral lines for T-Tauri stars often exhibit *P Cygni profiles,* as illustrated in Fig. 9.1, which indicate the presence of *expanding shells of low-density gas* around the stars.
- They are more luminous than corresponding main-sequence stars, implying that they are larger.
- They exhibit strong winds (*T-Tauri winds*), often with bipolar jet outflows having velocities of 300–400 km s^{-1}.

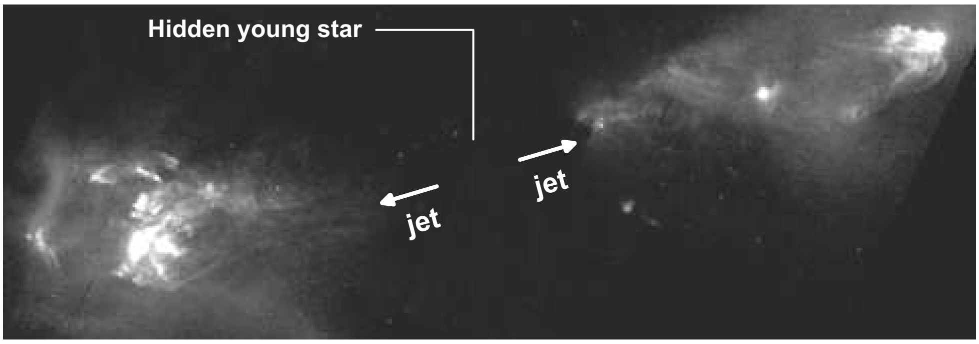


Figure 9.2: Jets and Herbig–Haro objects associated with outflow from young stars near the Orion Nebula. In the top image, the star responsible for the jets is hidden in the dark dust cloud lying in the center of the image. The entire width of this image is about one light year. The Herbig–Haro objects are designated HH-1 and HH-2, and correspond to the nebulosity at the ends of the jets. In the bottom image, a complex jet about a half light year long emerges from a star hidden in a dust cloud near the left edge of the image. The twisted nature of the jet suggests that the star emitting it is wobbling on its rotation axis, perhaps because of interaction with another star. The Herbig-Haro object HH-47 is the nebulosity on the right of the image. It is about 1500 light years away, lying at the edge of the Gum Nebula, which may be an ancient supernova remnant.

- *Herbig–Haro Objects* are often found in the directions of these jets.
- Two examples of outflow from young stars and associated Herbig–Haro objects are shown in Fig. 9.2.

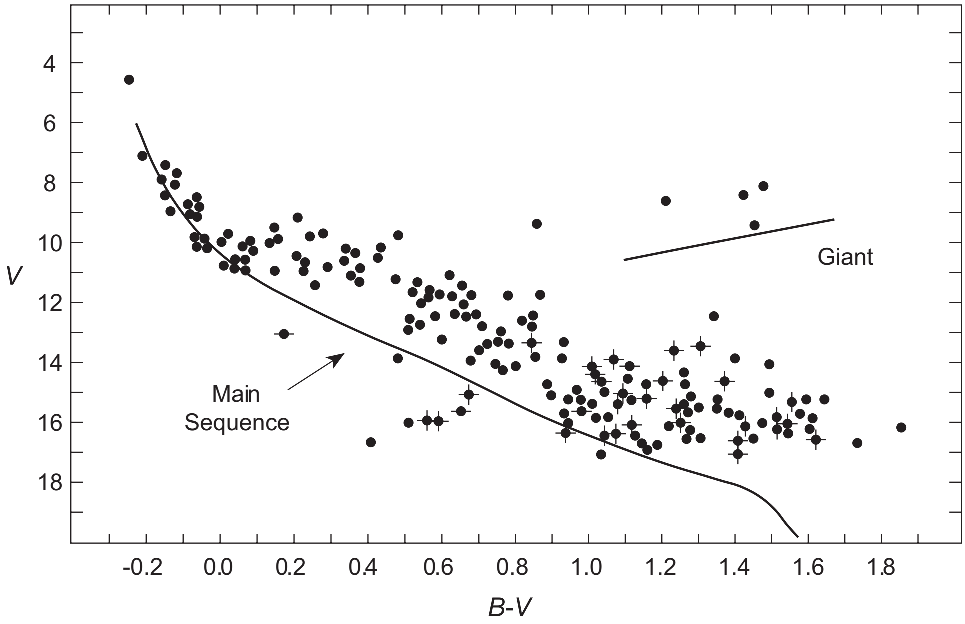


Figure 9.3: HR diagram for the young open cluster NGC2264. Horizontal bars denote stars with H_α line emission; vertical bars denote variable stars.

These considerations indicate that T-Tauri stars are *still in the process of contracting to the main sequence.*

- They are less massive than the O and B stars that often accompany them.
- Hence they will have contracted more slowly and many will not yet have had time to reach the main sequence.
- The HR diagram for a young cluster is illustrated in Fig. 9.3, where we see many young stars that have not yet reached the main sequence.
- Stars marked with horizontal and vertical bars in this figure have observational properties of T-Tauri stars.

- The bipolar outflows could in principle be explained by an *accretion disk* around the young T-Tauri stars.
- These would form as a result of *conservation of angular momentum* for the infalling matter.
- Then, if there are strong winds emanating from the star, they would tend to be directed in *bipolar flows* perpendicular to the plane of the accretion disk.
- However, it is difficult to explain
  - the *tight collimation of the jets* (as good as 10% over one parsec) by such a mechanism, and
  - the *energy driving the winds*

  in such a simple model.
- The *Herbig–Haro objects* are likely the result of
  - shocks formed when gas flowing out of the T-Tauri star interacts with clumps of matter, or when
  - clumps of matter ejected from the star interact with low density gas clouds.
- These observations suggest that
  - we must *look to the nebulae* to produce the stars and
  - the life of protostars contracting to the main sequence may be *more complex* than simple considerations would leave us to believe.

9.2 Conditions for Gravitational Collapse

Let's investigate the general question of *gravitational collapse to form stars* by considering a spherical cloud that

- is composed primarily of hydrogen,
- has a
  - radius R,
  - mass M, and
  - uniform temperature T, and
- consists of N particles of average mass μ, so that

$$M = N\mu M_u.$$

We shall assume that the question of stability is one of competition between

- *gravitation*, which would collapse the cloud, and
- *gas pressure*, which would expand the cloud.

9.2.1 The Jeans Mass and Jeans Length

- The gravitational energy is of the form

  $$\Omega = -f\frac{GM^2}{R},$$

  where the factor f

  - is *of order one* and
  - equal to $\frac{3}{5}$ if the cloud is spherical and of uniform density,
  - larger if the density increases toward the center.

- We take the thermal energy to be that of an ideal gas,

  $$U = \frac{3}{2}NkT.$$

- From the *virial theorem*, the static condition for gravitational instability is

  $$2U < |\Omega|,$$

  implying that the system is unstable if the mass M satisfies

  $$M > M_\mathrm{J} \equiv \frac{3kT}{fG\mu M_u}R = \left(\frac{3kT}{fG\mu M_u}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2},$$

  where

  $$N = \frac{M}{\mu M_u} \qquad R = \left(\frac{3M}{4\pi\rho}\right)^{1/3}$$

  have been employed.

- The *Jeans mass*

$$M_{\mathrm{J}} = \frac{3kT}{fG\mu M_u}R = \left(\frac{3kT}{fG\mu m_{\mathrm{H}}}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}$$

  appearing in the instability condition

$$M > M_{\mathrm{J}},$$

  defines a *critical mass for gravitational instability*.

- Since the *Jeans mass*
  - is proportional to $T^{3/2}\rho^{-1/2}$,
  - it will be *smaller for colder, denser clouds.*

> This makes physical sense: it is easier to collapse a cloud of a given mass gravitationally if the cloud is cold and dense than if it warm and diffuse.

- We may also solve for the *Jeans length*,

$$R_{\mathrm{J}} = \frac{fG\mu m_{\mathrm{H}}}{3kT}M_{\mathrm{J}}.$$

- The Jeans length R_{J}
  - defines the *characteristic length scale* associated with the Jeans mass and
  - characterizes the *minimum size of gravitationally unstable regions*.

9.2.2 The Jeans Density

- It is often more useful to express the Jeans criterion in terms of a critical density (the *Jeans density)*

$$\rho_{\mathrm{J}} = \frac{3}{4\pi M^2}\left(\frac{3kT}{f\mu m_{\mathrm{H}}G}\right)^3 .$$

- The critical density is lowest (thus easier to achieve) if
  - the *mass is large* and
  - the *temperature low*,

  as we would expect on intuitive grounds.

Example: Consider a cold cloud of molecular hydrogen, with

- $T = 20\,\mathrm{K}$
- $M = 1000 M_\odot$.

The associated Jeans density is only

$$\rho_{\mathrm{J}} = 10^{-22}\,\mathrm{g\,cm^{-3}}.$$

But a molecular hydrogen cloud at the same temperature with $M = 1\,M_\odot$ has a Jeans density that is 6 orders of magnitude larger.

- The Jeans criterion is *simple* because
  - it is a *static condition* that says nothing about gas dynamics and
  - it *neglects important factors* influencing stability such as
    - magnetic fields,
    - dust formation and vaporization, and
    - radiation transport.
- Nevertheless, the Jeans criterion is *a useful starting point* for understanding how stars form from clouds of gas and dust.

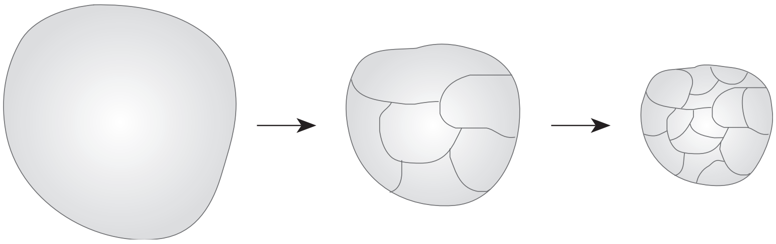


Figure 9.4: Fragmentation into gravitationally unstable subclouds.

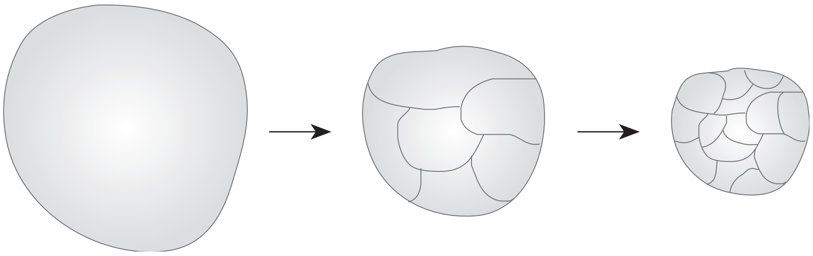
9.3 Fragmentation of Collapsing Clouds

From the foregoing, collapse of more massive clouds is favored, but most stars contain less than $1M_\odot$ of material.

- The solution to this dilemma is thought to lie in *fragmentation*, as illustrated in Fig. 9.4.
- As we shall see, the initial collapse is expected to occur at almost *constant temperature*. Therefore, from

$$M_{\mathrm{J}} = \left(\frac{3kT}{fG\mu m_{\mathrm{H}}}\right)^{3/2}\left(\frac{3}{4\pi\rho}\right)^{1/2}$$

  the *Jeans mass decreases with collapse* (ρ increases).
- *We speak loosely:* The Jeans criterion assumes a cloud near equilibrium, not one already collapsing.
- As a large cloud (small Jeans density) begins to collapse,
  - its average density increases with the collapse and
  - at some point subregions of the original cloud may exceed the critical density and become unstable in their own right toward collapse.



- If there are sufficient perturbations present in the cloud, these *subregions may separate and pursue independent collapse*.
- Within these subclouds the same sequence may be repeated: as the density increases, *subregions may themselves become gravitationally unstable* and begin an independent collapse.
- By such a *hierarchy of fragmentations*, it is plausible that clusters of protostars might be formed that have individual masses comparable to that of observed stars

9.4 Stability in Adiabatic Approximation

To understand further the behavior of gravitationally unstable clouds, let us consider the *adiabatic contraction* (or expansion) of a homogenous cloud.

- Real clouds will exchange energy with their surroundings and so are not completely adiabatic.
- However the results obtained in this limit will often be instructive in understanding more realistic situations.

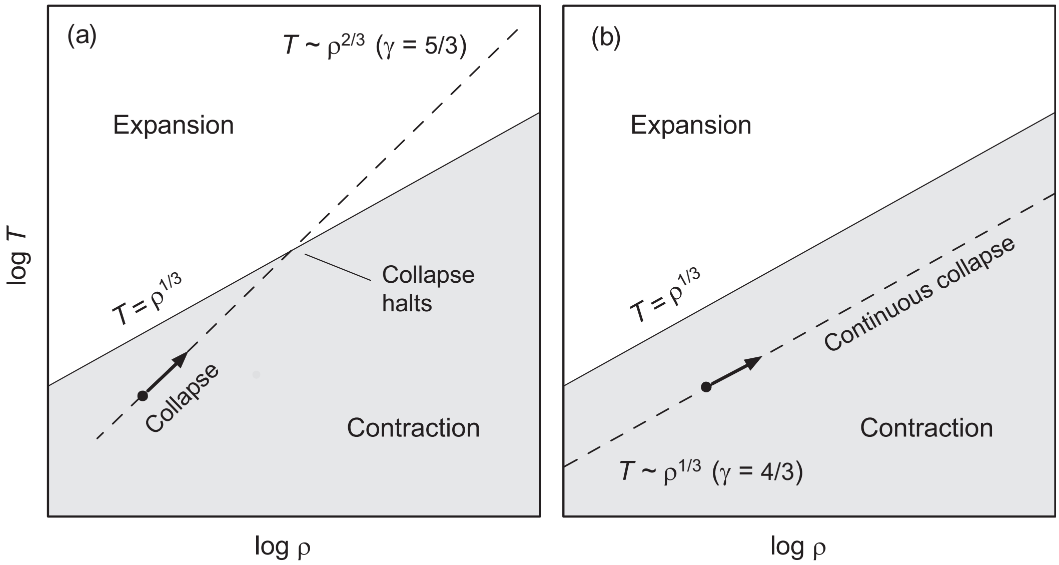


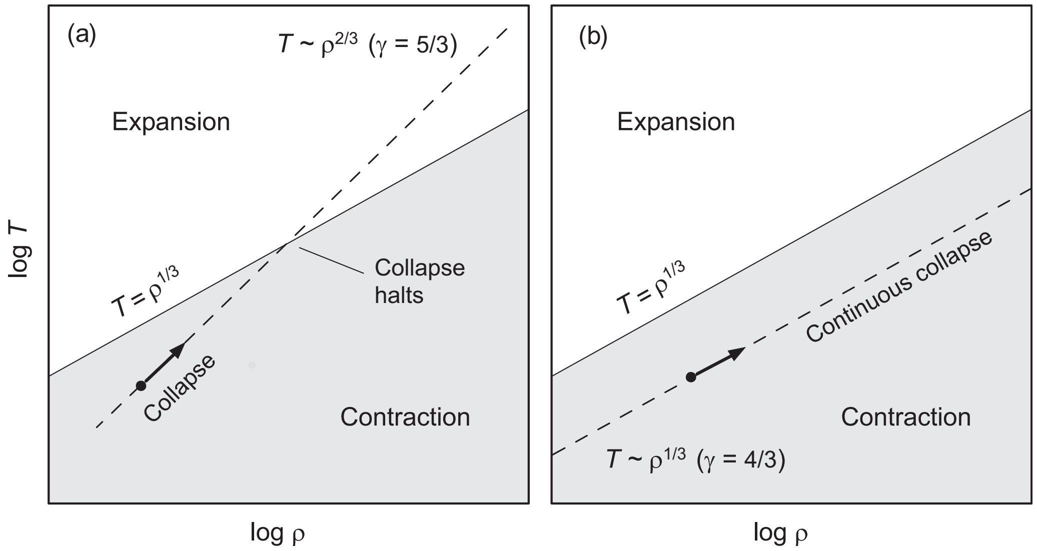
Figure 9.5: Gravitational equilibrium in temperature–density space.

- From

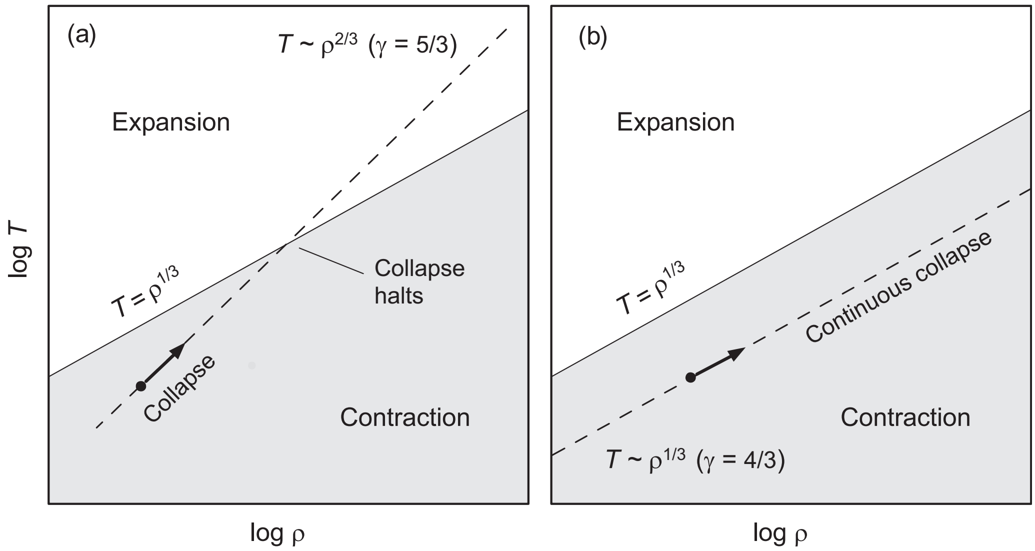
$$\rho_{\rm J} = \frac{3}{4\pi M^2}\left(\frac{3kT}{f\mu M_u G}\right)^3 .$$

  equilibration of gravity and pressure requires the temperature T and density ρ be related by $T \propto \rho^{1/3}$.

- In Fig. 9.5, this divides the T–ρ plane into
  - A region above the line $T \sim \rho^{1/3}$ where the system is unstable toward expansion, and
  - a region below the line where the system is unstable toward contraction.



- For points *above the stability line* (in the unshaded area), pressure forces are larger than the corresponding gravitational forces and the system is *unstable to expansion*.
- For points *below the stability line* (in the shaded area), pressure forces are weaker than the corresponding gravitational forces and the system is *unstable with respect to contraction*.

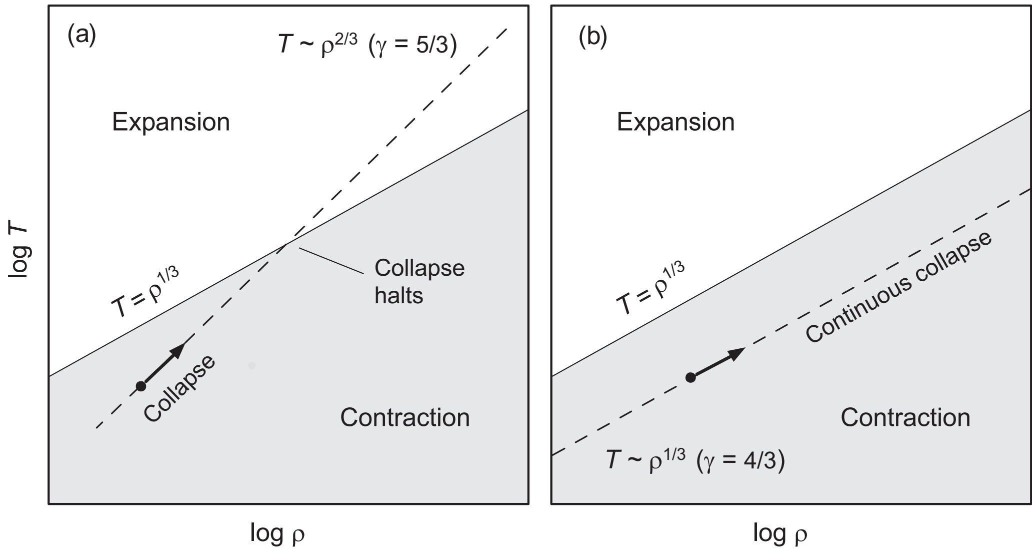


9.4.1 Dependence of Stability on Adiabatic Exponents

- First consider a *monatomic ideal gas*, for which *the adiabatic exponent is* $\gamma = \frac{5}{3}$.
- Since $\rho \propto V^{-1}$ and an adiabatic equation of state is $TV^{\gamma-1} =$ constant, for adiabats (paths followed in $\rho - T$ plane by adiabatic processes),

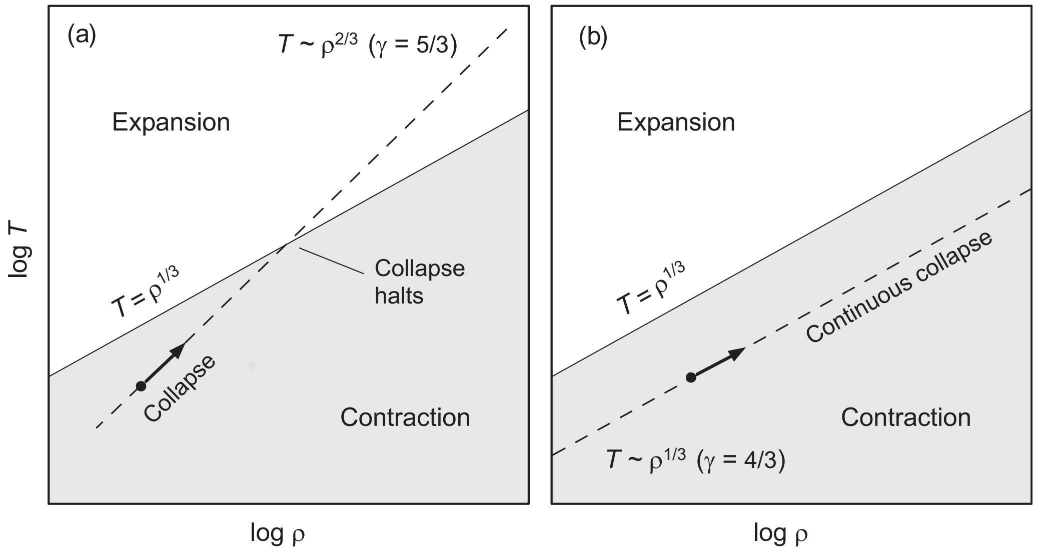
$$T \propto \rho^{\gamma-1} \quad \rightarrow \quad T\left(\gamma = \frac{5}{3}\right) = \rho^{2/3}.$$

- This corresponds to the dashed line in the left figure above, which is *steeper than the equilibrium line and therefore crosses it.*



- A cloud *unstable to contraction* corresponds to a point on the dashed line in the shaded area of the left figure.
- It will follow the dashed line to the right as it collapses, as indicated by the arrow *(right is increasing density).*
- The collapse will halt when the dashed line reaches the stability line (point labeled *"Collapse halts"*).
- Likewise, a cloud unstable to expansion corresponds to a point on the dashed line lying in the unshaded area of the left figure.
- It will follow the dashed line to the left as it expands *(left is decreasing density).*
- This expansion halts at the stability line.

Thus, $\gamma = \frac{5}{3}$ *is gravitationally stable.*



- Now consider the right figure above, where we assume that *the cloud has an adiabatic exponent* $\gamma = \frac{4}{3}$.
- In this case, the contraction (or expansion) follows an adiabat for which $T \propto \rho^{\gamma-1} \propto \rho^{1/3}$.
- Since this adiabat is *parallel to the stability line*,
  - *the two lines never cross* and
  - a system lying on the dashed line collapses and *continues to collapse* adiabatically.
- This will also be the case if $\gamma < \frac{4}{3}$.
- Likewise, a system with $\gamma = \frac{4}{3}$ that is above the stability line *expands adiabatically as long as* $\gamma = \frac{4}{3}$.

Thus, $\gamma \leq \frac{4}{3}$ is *gravitationally unstable.*

9.4.2 Physical Interpretation

Physical meaning of the preceding discussion:

- A gas is less able to generate the *pressure differences required to resist gravity* if the energy released by gravitational contraction can be absorbed into *internal degrees of freedom*.
- This energy is *not available* to increase the kinetic energy of the gas particles.
- The parameter γ is relevant because it is related to the heat capacities for the gas ($\gamma = C_\text{P}/C_\text{V}$ for ideal gas), which depends on the number of degrees of freedom for particles in the gas.
- Typical sinks of energy that can siphon off energy internally are

  1. rotations of molecules,
  2. vibrations of molecules,
  3. ionization, and
  4. molecular dissociation.

Such internal degrees of freedom are *energy sinks that lower the resistance of the gas to gravitational compression*.

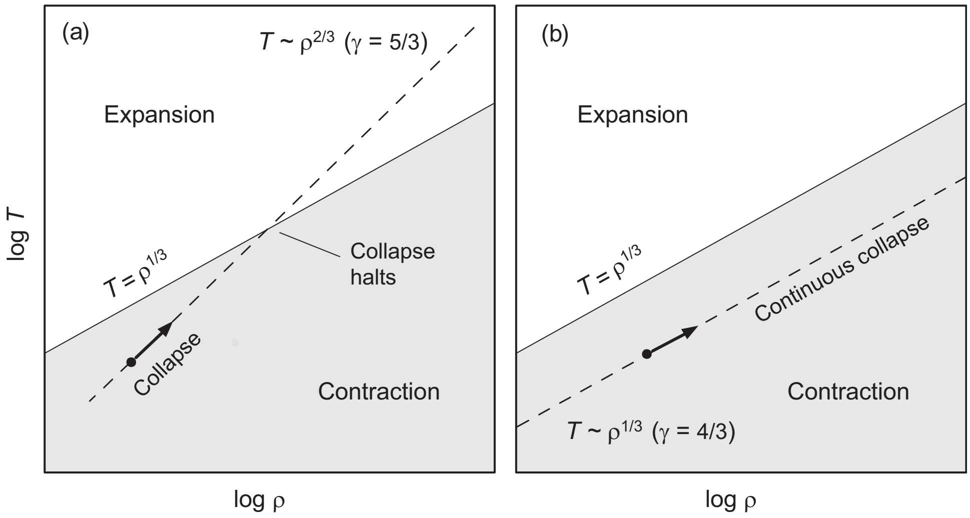
- In large clouds γ can be reduced to $\frac{4}{3} = 1.33$ or less by
  1. *Polyatomic gases* with $s > 5$,

$$\gamma = \frac{1+s/2}{s/2} \quad \rightarrow \quad \gamma(s=5) = \frac{1+5/2}{5/2} = \frac{7}{5} = 1.4,$$

  2. *Ionization of hydrogen* around $10,000\,\mathrm{K}$.
  3. *Dissociation of hydrogen molecules* around $4,000\,\mathrm{K}$.

- The large molecules required for the first situation are *relatively rare in the interstellar medium but very effective*.

- In hydrogen ionization or molecular dissociation zones,
  - Typically $\gamma \leq \frac{4}{3}$ and this causes an instability until the ionization or dissociation is complete.
  - Then γ will return to normal values ($\gamma \simeq \frac{5}{3}$) and collapse on the corresponding adiabat will reach the equilibrium line and stabilize the collapse.



9.5 The Collapse of a Protostar

The preceding introduction sweeps much under the rug.

- But we shall assume that the existence of stars implies that protostars form by some mechanism similar to the one outlined above.
- Let us consider the *collapse of a one solar mass protostar*.
- From the *Jeans criterion* for $T = 20\,\text{K}$ and $M = 1M_\odot$,

$$\rho_\text{J} \simeq 3 \times 10^{-16}\,\text{g cm}^{-3}.$$

- Thus, we expect that a $1M_\odot$ cloud can collapse if this average density is exceeded.
- The size of this initial cloud may be estimated by assuming the density to be
  - constant and
  - distributed spherically

  implying that $R \sim 3 \times 10^{16}\,\text{cm} \sim 2000\,\text{AU}$.
- Thus, the initial protostar has a radius approximately *50 times that of the present Solar System*.

9.5.1 Initial Free-Fall Collapse

- The initial collapse is *free-fall* and *isothermal*, as long as the gravitational energy released is not converted into thermal motion of the gas and thereby into pressure.

- This will be the case as long as the energy not radiated away is largely taken up by

  1. *dissociation of hydrogen molecules* into hydrogen atoms
  2. *ionization* of the hydrogen atoms.

- The *dissociation energy* for hydrogen molecules is $\varepsilon_{\mathrm{d}} = 4.5\,\mathrm{eV}$

- The *ionization energy* for hydrogen atoms is $\varepsilon_{\mathrm{ion}} = 13.6\,\mathrm{eV}$.

- The *energy required to dissociate and ionize all the hydrogen* in the original cloud is then

  $$\begin{aligned} E &= N(H_2)\varepsilon_{\mathrm{d}} + N(H)\varepsilon_{\mathrm{ion}} \\ &= \frac{M}{2m_{\mathrm{H}}}\varepsilon_{\mathrm{d}} + \frac{M}{m_{\mathrm{H}}}\varepsilon_{\mathrm{ion}}, \end{aligned}$$

  where N denotes the number of the corresponding species and m_{H} is the mass of a hydrogen atom.

- For the case of a protostar of one solar mass, the requisite energy is approximately 3×10^{46} erg.

- If the dissociation and ionization energy

$$E = \frac{M}{2m_{\rm H}}\varepsilon_{\rm d} + \frac{M}{m_{\rm H}}\varepsilon_{\rm ion},$$

  is supplied by contraction from radius R_1 to R_2,

$$\underbrace{\frac{GM^2}{R_2} - \frac{GM^2}{R_1}}_{\text{gravity}} = \underbrace{\frac{M}{2m_{\rm H}}\varepsilon_{\rm d} + \frac{M}{m_{\rm H}}\varepsilon_{\rm ion}}_{\text{dissociation and ionization}} .$$

- Solve for R_2 with $M = 1M_\odot$ and $R_1 = 3 \times 10^{16}\,$cm to give

$$R_2 = 9 \times 10^{12}\,\text{cm} \simeq 130R_\odot \simeq 0.6\,\text{AU}.$$

- The corresponding *dynamical timescale* is

$$t_{\rm ff} = \sqrt{\frac{3\pi}{32G\rho}} \simeq 13{,}000\,\text{yr}.$$

Thus a $1\,M_\odot$ protostar collapses in *near free-fall*

- from about 50 times the radius of the Solar System
- to about half the radius of the Earth's orbit

in $\sim 10^4$ years.

- The collapse then slows because

  1. All the hydrogen has been dissociated and ionized,
  2. the photon mean free path becomes short and the cloud becomes opaque to its own radiation,
  3. temperature increases as heat is trapped, and
  4. pressure gradients counteract gravity and bring the system into near hydrostatic equilibrium.

Thus, we may apply the *virial theorem in near adiabatic conditions* from this point onward.

9.5.2 Homology

From the expressions for free-fall collapse we see that the characteristic timescale for free fall is independent of the radius of the collapsing mass distribution.

- This behavior is termed *homologous collapse.*
- One consequence of homologous collapse is that if the initial density is uniform it remains uniform for the entire collapse.
- Because successive configurations in homologous processes are self-similar (related by a scale transformation), homologous systems are particularly simple to deal with mathematically.
- Therefore, reasonably good approximate treatments of the initial phases of gravitational collapse are often possible by making homology assumptions.

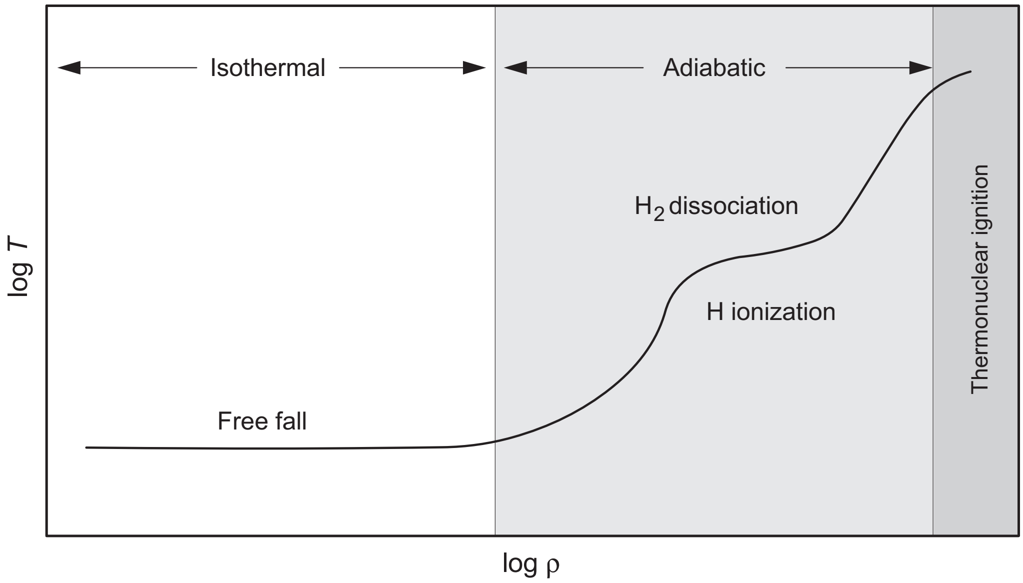


Figure 9.6: Schematic track for the collapse of a gas cloud to form a star.

9.5.3 A More Realistic Picture

The preceding picture is *oversimplified*. A more realistic variation of temperature and density for star formation is illustrated in Fig. 9.6.

- In this more realistic picture the cloud begins to heat and deviate from free-fall once it traps significant heat.
- When the temperature is sufficient to dissociate and then ionize hydrogen, the cloud again collapses $\sim$ isothermally for a time.
- Once all hydrogen has been dissociated and ionized, the collapse returns to one governed by approximately adiabatic conditions in near hydrostatic equilibrium.

9.6 Onset of Hydrostatic Equilibrium

The temperature at which hydrostatic equilibrium sets in may be estimated as follows.

- From the virial theorem we have that $2U + \Omega = 0$.
- The gravitational energy Ω is

  $$\Omega \equiv \Omega(R_2) = -\frac{GM^2}{R_2} = -\left(\frac{M}{2m_{\mathrm{H}}}\varepsilon_{\mathrm{d}} + \frac{M}{m_{\mathrm{H}}}\varepsilon_{\mathrm{ion}}\right),$$

  where $\Omega(R_1)$ has been neglected compared with $\Omega(R_2)$.
- From $U = \frac{3}{2}NkT$, the internal energy for the hydrogen ions and electrons in the fully ionized gas is approximately

  $$U \simeq \frac{3}{2}(N_{\mathrm{H}} + N_{\mathrm{e}})kT = 3N_{\mathrm{H}}kT = \frac{3M}{m_{\mathrm{H}}}kT,$$

  where N_{H} is the number of hydrogen ions and $N_{\mathrm{e}} \sim N_{\mathrm{H}}$ is the number of free electrons.
- Therefore, the virial theorem requires that

  $$2U + \Omega = \frac{6M}{m_{\mathrm{H}}}kT - \frac{M}{2m_{\mathrm{H}}}\varepsilon_{\mathrm{d}} - \frac{M}{m_{\mathrm{H}}}\varepsilon_{\mathrm{ion}} = 0,$$

  and solving this for T gives

  $$T = \frac{1}{k}\left(\frac{\varepsilon_{\mathrm{d}}}{12} + \frac{\varepsilon_{\mathrm{ion}}}{6}\right) \simeq \frac{2.6\,\mathrm{eV}}{k} \simeq 30,000\,\mathrm{K}$$

  for the onset of hydrostatic equilibrium.

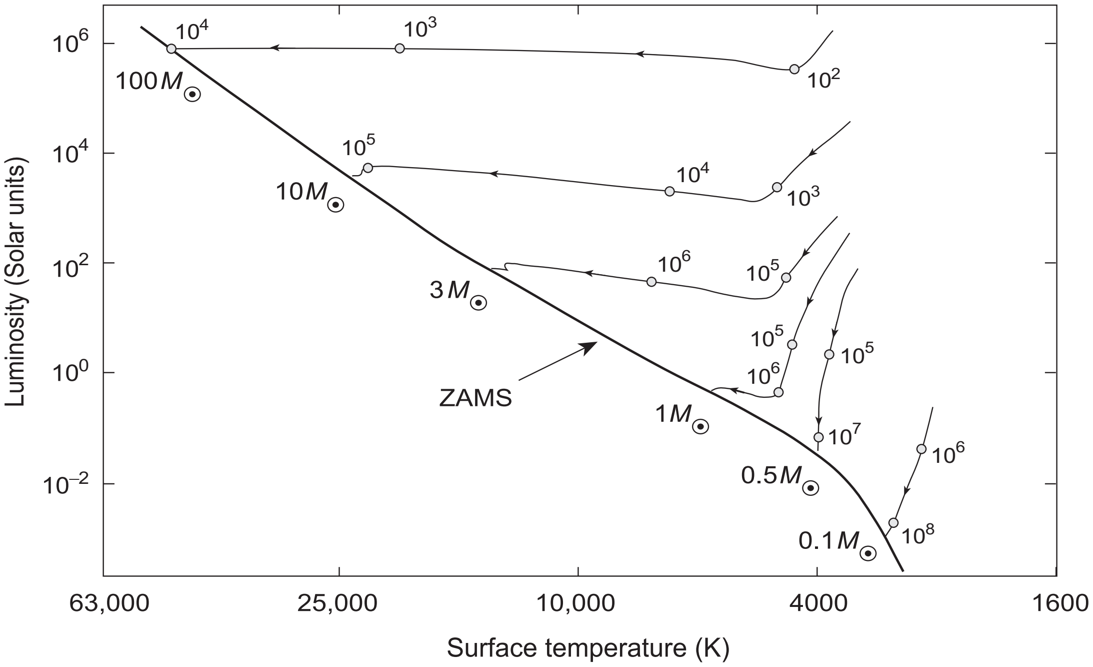


Figure 9.7: Evolutionary tracks for collapse to the main sequence. Numbers on tracks are times in years.

- Subsequent contraction of the protostar is
  - in near hydrostatic equilibrium and
  - is *controlled by the opacities*,

  which govern how fast energy can be brought to the surface and radiated.
- This too is a consequence of the virial theorem and leads to the *Kelvin–Helmholtz timescale* of about 10^7 years for a star of one solar mass.
- The evolutionary tracks for protostars of various masses to collapse to the main sequence are shown in Fig. 9.7.

9.7 Termination of Fragmentation

- Earlier we indicated that collapse of large clouds is likely to fragment into a hierarchy of sub-collapses, explaining why we observe many low-mass stars.
- However, this argument is incomplete: we must ask what stops the fragmentation in the vicinity of $0.1 - 1\,M_\odot$.
- The likely answer is that *the transition from isothermal to adiabatic collapse implies a modification of the Jeans criterion* and that this dictates a lower limit for the mass of the fragments produced by a hierarchical collapse.
- Substitution of the condition $T \propto \rho^{\gamma-1}$ for adiabats in

  $$M_\mathrm{J} = \left(\frac{3kT}{fG\mu m_\mathrm{H}}\right)^{3/2} \left(\frac{3}{4\pi\rho}\right)^{1/2},$$

  implies that for adiabatic clouds

  $$M_\mathrm{J} \simeq \rho^{(3\gamma-4)/2}.$$

  For $\gamma = 5/3$ then, the Jeans mass is proportional to $\rho^{1/2}$ and *in adiabatic collapse the Jeans mass increases.*
- This implies that *the transition from isothermal to adiabatic collapse sets a lower bound on possible Jeans masses.*

- More realistic calculations do suggest a lower bound to masses that can collapse gravitationally that is controlled by cloud opacities.
- However, this bound is often lower than the $M \sim 0.1 M_\odot$ found for the lightest stars.
- However, we shall see later that the lightest fragments with mass less than $M \sim 0.1 M_\odot$ can collapse to *brown dwarfs*, which
  - form by gravitational collapse but
  - are not stars.

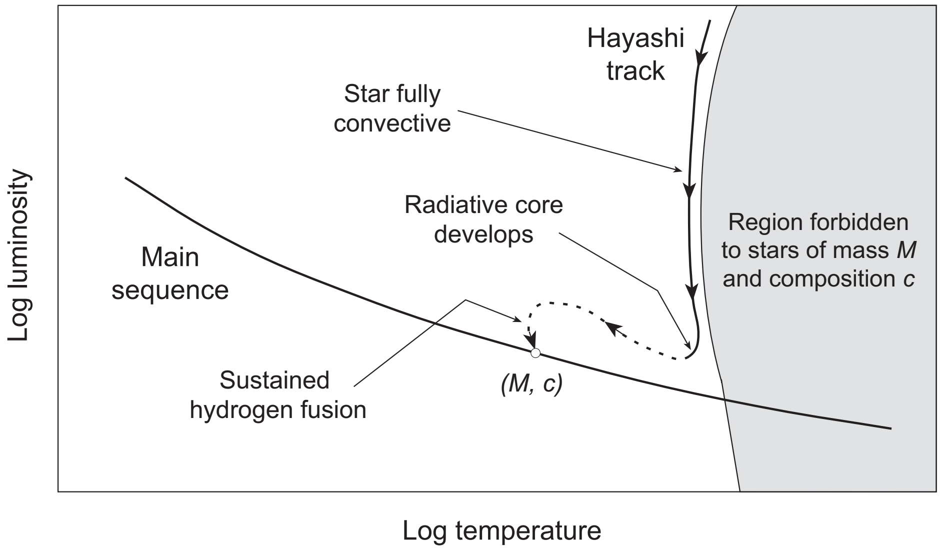


Figure 9.8: Evolution of protostars to the main sequence.

9.8 Hayashi Tracks

A deeper understanding of the collapse to the main sequence follows from a fundamental result first obtained by Hayashi:

> A star generally cannot reach hydrostatic equilibrium if its surface is too cool.

- This implies a region in the HR diagram that is forbidden to a given star if it is in hydrostatic equilibrium.
- This region is called the *Hayashi forbidden zone;* it is illustrated in Fig. 9.8 for a star of mass M and composition c.

9.8.1 Fully Convective Stars

Stars contracting to the main sequence

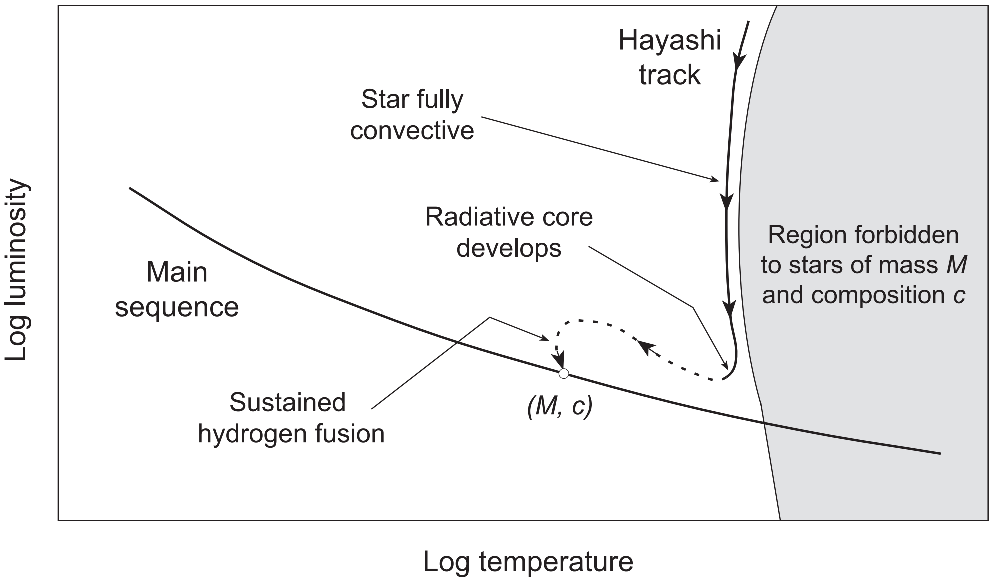
- Must have large surface areas and (relatively) high surface temperatures, so they have *large luminosities.*
- Once hydrogen is ionized they have *high opacities.*
- The combination of
  - *high opacity* with
  - *large luminosity*

  ensures that the *temperature gradients exceed the adiabatic one.*
- Thus *such forming stars are almost completely convective*.
- Recall that completely convective stars are approximately described by a $\gamma = \frac{5}{3}$ polytrope:

For a completely ionized star, fully mixed by convection with negligible radiation pressure, if $\gamma = \frac{5}{3}$,

$$P = K\rho^{\gamma}(r) = K\rho^{1+1/n}(r) = K\rho^{5/3},$$

- where $n = 1/(\gamma - 1)$,
- and where K is constant for a given star.



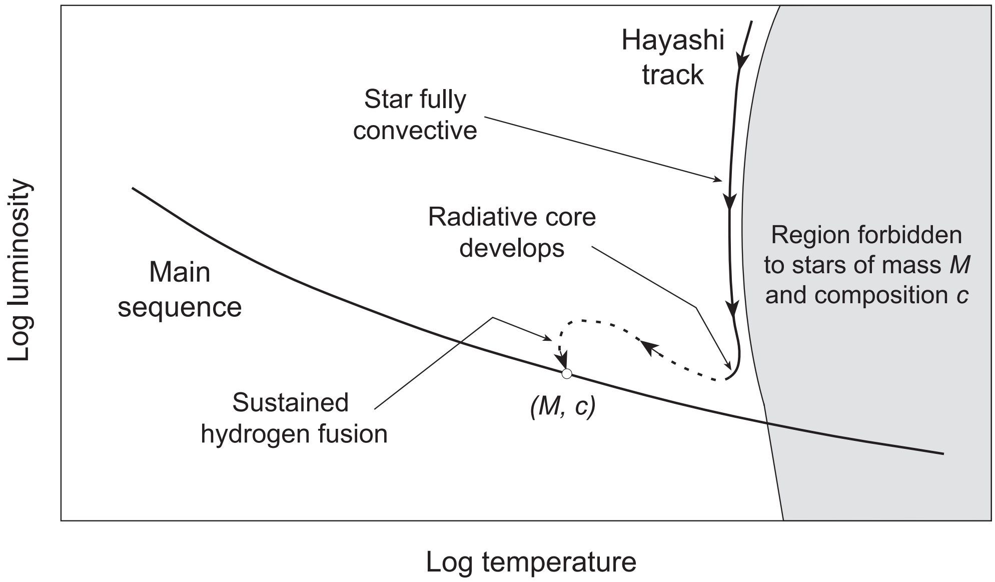
By examining *fully convective stars* with a *thin radiative envelope*, Hayashi showed that

- Contracting protostars follow an almost vertical HR trajectory called the *Hayashi track* for the star.
- If the star is fully convective, the Hayashi track is essentially defined by the left boundary of the Hayashi forbidden zone, as illustrated in the above figure.
- Numerical simulations and simplified models indicate that objects to the right of the Hayashi track *cannot achieve hydrostatic equilibrium*.
- Thus no stable protostars can exist the the forbidden region.

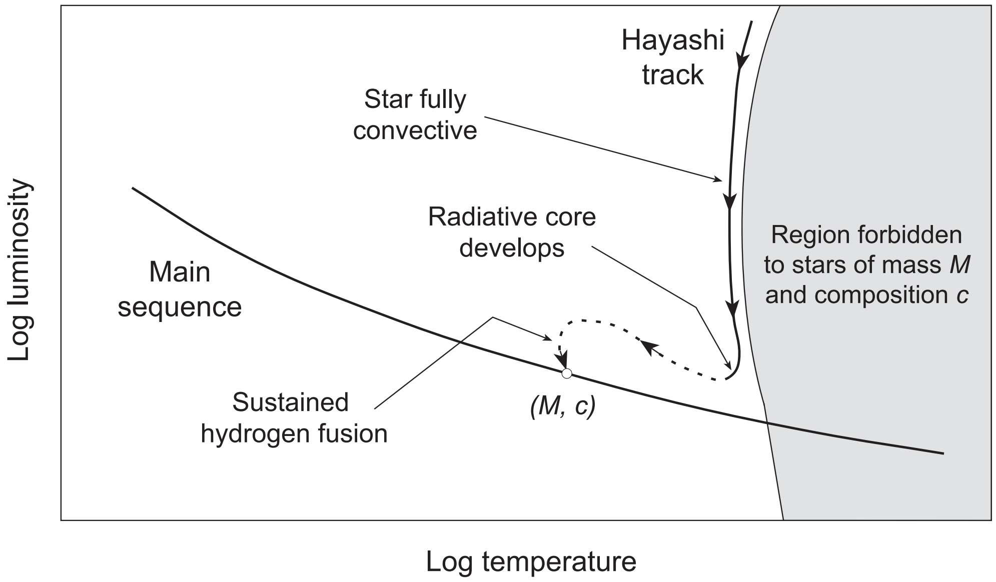
9.8.2 Development of a Radiative Core

As the collapsing star descends the Hayashi track its central temperature is increased by the gravitational contraction.

- This *decreases the central opacity* (recall that for the representative Kramers opacity, $\kappa \sim \rho T^{-3.5}$).
- Eventually this lowers the temperature gradient in the central region sufficiently that it drops below the critical value for convective stability.
- A *radiative core* develops.
- As contraction proceeds the radiative core begins to grow at the expense of the convective regions, which are eventually pushed out to the final subsurface regions characteristic of stars like the Sun.
- (In more massive stars the subsurface convective zones are eliminated completely but the core may become convective if the power generation is sufficiently large after the star enters the main sequence.)



- For the fully convective star on the Hayashi track, luminosity decreases rapidly (shrinking surface area).
- However, as the protostar shrinks in size,
  - *opacity decreases* over more and more of the interior because of the increasing temperature,
  - *luminosity begins to rise again* because more energy can flow out radiantly.
- Since at this point the star is *shrinking as its luminosity increases,* its *surface temperature must increase.*
- The star begins to follow a track to the left and somewhat upward in the HR diagram (above figure).
- Finally, *onset of hydrogen fusion* causes the track to bend over and enter the main sequence.



Thus, the contraction to the vicinity of the main sequence is composed of two basic periods:

1. A *vertical descent* in the HR diagram for fully convective stars, followed by
2. a *drift up and to the left* as the interior of the star becomes increasingly radiative at the expense of the convective envelope.

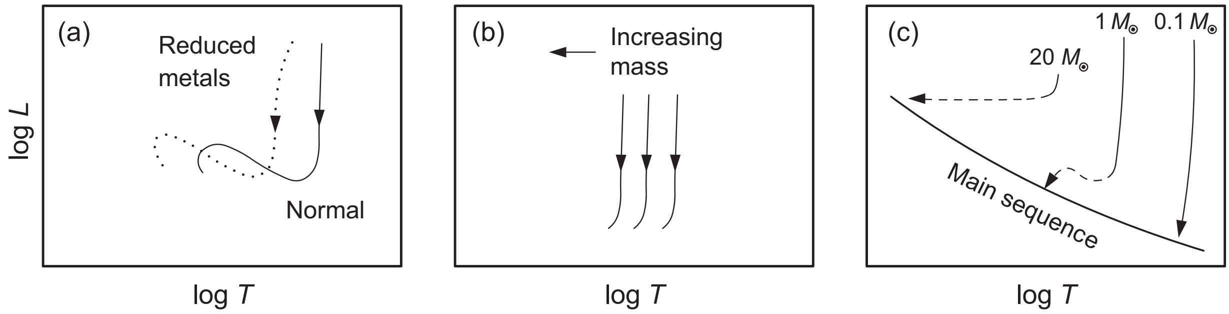


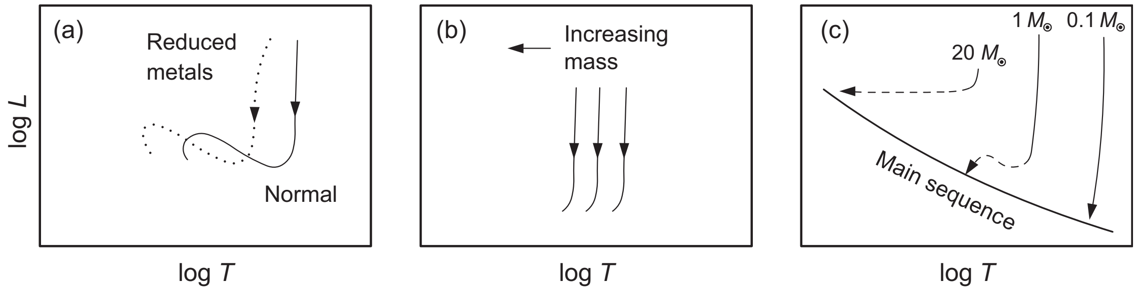
Figure 9.9: Dependence of Hayashi tracks on (a) composition and (b) mass. The solid portions of each curve in (c) represent the descent on the Hayashi track.

9.8.3 Dependence of Track on Composition and Mass

Hayashi tracks *depend weakly on the mass and composition,* as illustrated in Fig. 9.9.

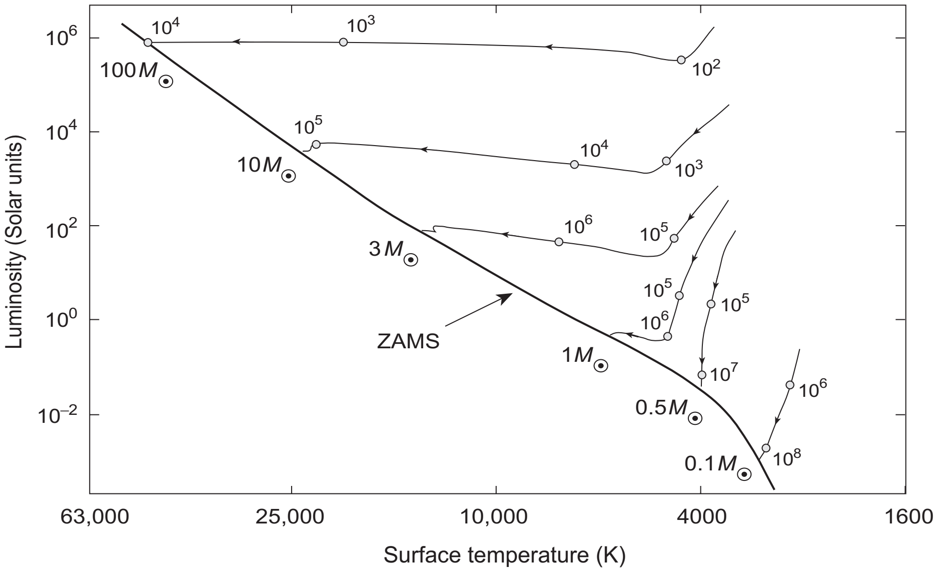
- For more massive stars of fixed composition
  - the Hayashi tracks are almost parallel to each other, but
  - are increasingly shifted to the left in the HR diagram

  (see Fig. 9.9b).
- The Hayashi tracks also depend on *stellar composition,* because this can influence the *opacity.*
- For example, a metal-poor star of a given mass will generally have a Hayashi track to the left of an equivalent metal-rich star because of lower opacity (Fig. 9.9a).



- The transition from convective to radiative interiors, and the corresponding transition from downward to more horizontal leftward HR motion, is *faster in more massive stars because of more rapid interior heating.*

- As illustrated in Fig. (c) above and the following figure, *massive stars leave the Hayashi track quickly* and approach the main sequence almost horizontally.



- Conversely, *the least massive stars never leave the Hayashi track* and are thought to *remain completely convective,* even after entering the main sequence.

9.9 Limiting Lower Mass for Stars

A contracting protostar will become a star only if the temperature increases sufficiently in the core to initiate thermonuclear reactions.

- For an idealized star composed of a monatomic ideal gas having uniform temperature and density, the temperature varies with the cube root of the density:

$$T = 4.09 \times 10^6 \mu \left(\frac{M}{M_\odot} \right)^{2/3} \rho^{1/3}.$$

- However, this behavior assumes an ideal classical gas; *the temperature will no longer increase with contraction if the equation of state becomes that of a degenerate gas.*

- The critical temperature and density for onset of electron degeneracy can be estimated by *setting kT equal to the fermi energy*, which gives a critical density

$$\rho \simeq 6 \times 10^{-9} \mu_e T^{3/2} \,\mathrm{g\,cm^{-3}}.$$

- Inserting this into the preceding equation gives for the temperature at which the critical density is reached in the contracting protostar,

$$T \simeq 5.6 \times 10^7 \mu \mu_e^{1/3} \left(\frac{M}{M_\odot} \right)^{2/3}.$$

- For $M \sim M_\odot$ and $\mu \mu_e^{1/3} \sim 1$, we obtain $T \sim 10^7$ K from

$$T \simeq 5.6 \times 10^7 \mu \mu_e^{1/3} \left(\frac{M}{M_\odot} \right)^{2/3} .$$

- Thus, *a solar mass protostar can produce an average temperature of 10 million K by contraction.*
- This is more than enough to ignite hydrogen fusion before the electrons in the core become degenerate .
- On the other hand, as the mass of the protostar is decreased we will eventually reach a mass where *the core will become degenerate before the temperature rises to the hydrogen fusion temperature.*
- Detailed calculations indicate that this limiting mass is approximately $0.08 M_\odot - 0.10 M_\odot$.
- What of collapsing clouds with less than this critical mass required to form stars?
- For them the growth in temperature is halted by electron degeneracy pressure before fusion reactions can begin and *no star is formed.*
- It is speculated that many such objects may exist,
  - supported hydrostatically by electron degeneracy
  - radiating energy left over from earlier contraction.

  Such objects are called *brown dwarfs.*

9.10 Brown Dwarfs

Brown dwarfs collapse out of hydrogen clouds, not out of protoplanetary disks (like stars).

- But they radiate energy only by gravitational contraction, not from hydrogen fusion (like planets).
- Their masses are expected to range from several times the mass of Jupiter to a few percent of the Sun's mass.
- The cooler brown dwarfs may resemble gas giant planets in chemical composition,
- Hotter ones may begin to look chemically more like stars.
- They are difficult to detect since they are small and of very low luminosity.

9.10.1 Spectroscopic Signatures

The first brown dwarf discovered was Gliese 229B.

- Gliese 229B appears to be
  - too hot and massive to be a planet, but
  - too small and cool to be a star.
- The IR spectrum of GL229B looks like the spectrum of a gas giant planet.
- Most telling is *evidence of methane gas,*
  - which is common in gas giants but
  - is not found in stars

  because methane molecules can survive only in atmospheres having temperatures lower than about 1500 K.

In addition to searching for gases like methane that should not be present in stars, searches for brown dwarfs have also looked for *evidence of the element lithium.*

- Hydrogen fusion destroys lithium in stars.
- At temperatures above about 2×10^6 K, a proton encountering a lithium nucleus has a high probability to react with it, converting the lithium to helium.
- The amount of lithium that can survive is a function of how strongly the material of the star is mixed down to the core fusion region by convection.
- Protostars are convective, so stars start off with a strongly mixed interior, but the initial core temperature in the protostar is not high enough to burn lithium.
- The lightest stars (red dwarfs) remain convective once on the main sequence, so lithium is mixed down to the fusion region and destroyed in red dwarfs.
- Because these stars are cool, it takes some time to burn the lithium.
- Calculations indicate that *lithium could survive no longer than about* 2×10^8 *years in the lightest true star.*

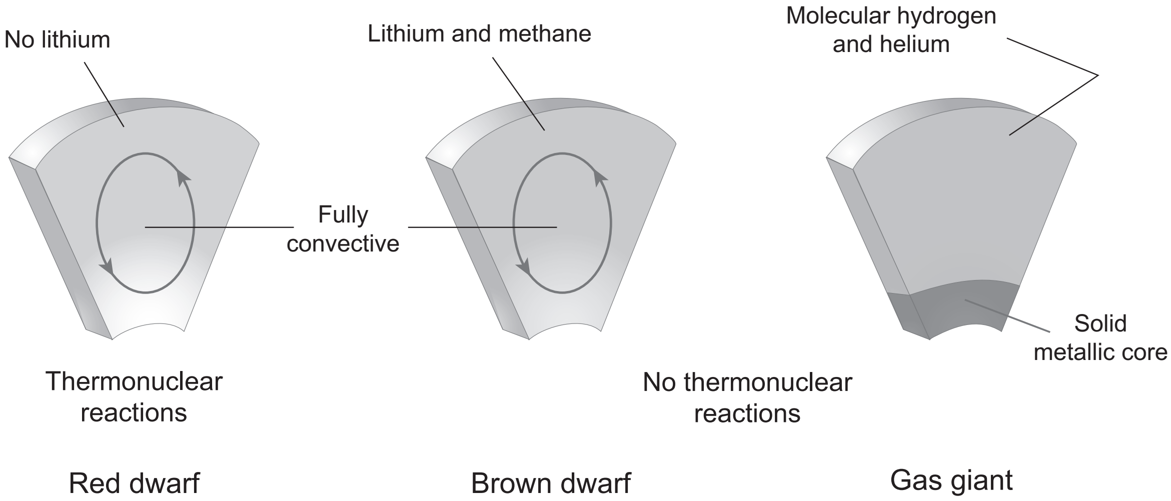


Figure 9.10: Contrasting interiors of a red dwarf, a brown dwarf, and a gas giant planet. Generally stars initiate thermonuclear reactions but brown dwarfs and planets do not. Thus, lithium is destroyed in stars. The presence of methane is also an indication that the temperatures are too low for the object to be a star. Gas giant planets can also contain lithium and methane, but their upper interiors tend to be dominated by molecular hydrogen and helium.

The basic interior structures expected for stars, brown dwarfs, and gas giant planets are summarized schematically in Fig. 9.10.

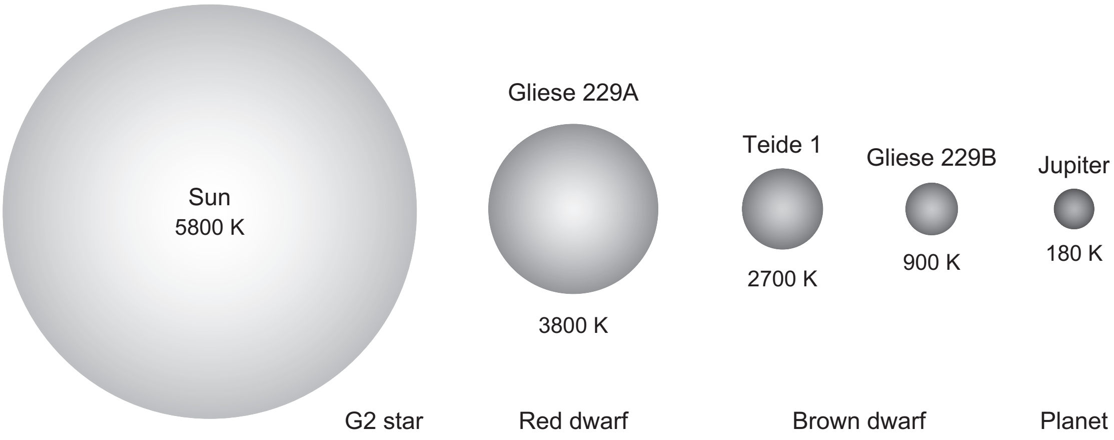


Figure 9.11: Size and surface temperature trends for stars, brown dwarfs, and gas giant planets.

9.10.2 Stars, Brown Dwarfs, and Planets

Figure 9.11 summarizes size and surface temperature trends from stars like the Sun, through the lowest mass stars (red dwarfs), brown dwarfs, and finally to planets.

- Brown dwarfs can have surface temperatures comparable to that of the lowest mass stars, but atmospheric compositions similar to large planets.
- The challenge is to distinguish brown dwarfs from stars and gas giants at interstellar distances.
- Many of brown dwarf candidates have been identified.
- However, in many cases there is uncertainty about whether they are brown dwarf companions of stars, or giant planets orbiting stars.

Why are Stars Hot?

Let us tie together threads involving hydrostatic equilibrium, the virial theorem, stellar energy production, and gravitational collapse by asking: *“Why are stars hot?”*

- The popular perception is that stellar cores are hot because they correspond to enormous thermonuclear furnaces.
- But the core of the Sun is in fact a very low density power source (*a few hundred watts per cubic meter in the core).*
- As we have seen, it is
  - release of gravitational energy by *contraction* and
  - partial trapping of that energy by *high stellar opacity*

  that raises the protostar interior to fusion temperature.
- *The role of hydrogen burning is not to heat stars* (gravity and the virial theorem can do that); it is to *sustain the luminosity* over much longer periods than would be possible otherwise.
- Triggering thermonuclear reactions replaces the *Kelvin–Helmholtz timescale* for sustained luminosity with the much longer *nuclear burning timescale.*
- This enables the Sun to radiate its present power for billions of years rather than millions of years.
- So the source of sustained luminosity and sustained high interior temperatures for main sequence stars is indeed fusion, but *the cores of those stars were heated by gravitational contraction,* to their present temperatures.
- They *maintain* those temperatures and luminosities by slow fusion reactions in hydrostatic equilibrium.

9.11 Limiting Upper Mass for Stars

- A limiting lower mass for stars is set by the requirement that sufficient temperature be generated by gravitational collapse to commence the burning of hydrogen to helium in the core.
- An upper limiting mass for stars is thought to exist because of the opposite extreme:

> *If a star is too massive, the intensity of the energy production makes the star unstable to disruption by the radiation pressure.*

- The *pressure associated with the radiation grows as* T^4 and thus will be most important for very hot stars.
- It is instructive to ask what photon luminosity is required such that the magnitude of the radiation force is equivalent to the magnitude of the gravitational force.
- This critical luminosity, which defines limiting configurations that are stable gravitationally with respect to the pressure of the photon flux, is termed the *Eddington luminosity*.

9.11.1 Eddington Luminosity

- The force per unit volume associated with a photon gas is given by the gradient of the radiation pressure,

$$\frac{1}{V}F_{\mathrm{r}} = -\frac{dP_{\mathrm{r}}}{dr} = \frac{4}{3}aT^3\frac{dT}{dr}.$$

- Equating the magnitudes of this force and the gravitational force gives an expression for the *Eddington luminosity*

$$L_{\mathrm{Edd}} = \frac{4\pi cGM}{\kappa}.$$

- We may expect that stars exceeding this luminosity can *blow off surface layers by radiation pressure*.
- The ejection of material
  - may also be aided by *stellar pulsations* that result from pressure instabilities at high luminosity, and
  - may be influenced by *rotation and magnetic fields*.
- Force from radiation pressure and from gravity are proportional to ρ/r^2, so the dependence on r and ρ cancels from the Eddington luminosity and

  1. the *mass* of the star and
  2. *opacities* for regions near the surface.

  determine completely the critical luminosity.

9.11.2 Estimate of Limiting Mass

The Eddington luminosity may be expressed as

$$\frac{L_{\text{Edd}}}{L_\odot} \simeq 3.5 \times 10^4 \left(\frac{M}{M_\odot}\right),$$

if we estimate the opacity κ by the *Thomson formula*.

- We may use this equation to estimate a radiation-pressure mass limit by
  - assuming that the most luminous stars observed ($L \sim$ several $\times\, 10^6 L_\odot$) are *at the Eddington limit*, and
  - approximating the relevant opacity by the Thomson formula.
- This suggests a maximum stable mass of order $100 M_\odot$.

This is a very crude estimate but calculations, and observations, suggest that the most massive stars indeed have masses of this order.

9.12 The Initial Mass Function

Mass is destiny for stars. The distribution in initial mass for a population of stars is called the *initial mass function (IMF)* for that population.

- The initial mass function $\xi(M)$ is defined by requiring that the mass bound up in stars in the interval M to $M+dM$ in a volume of space be given by

  $$MdN = \xi(M)dM,$$

  where N is the number of stars in the volume.

- Determining the IMF requires an indirect, semiempirical chain of reasoning since
  - *Observations give apparent magnitudes*, not masses.
  - *Stellar populations evolve*, so the present mass distribution differs from the initial one.

- Edwin Salpeter first estimated $\xi(M)$ in 1955 by
  - Examining the luminosities of main sequence stars in the neighborhood of the Sun.
  - Relating the luminosity to the mass by empirical mass–luminosity relations.
  - Assuming that stars evolve away from the main sequence when about 10% of their initial hydrogen has been burned.

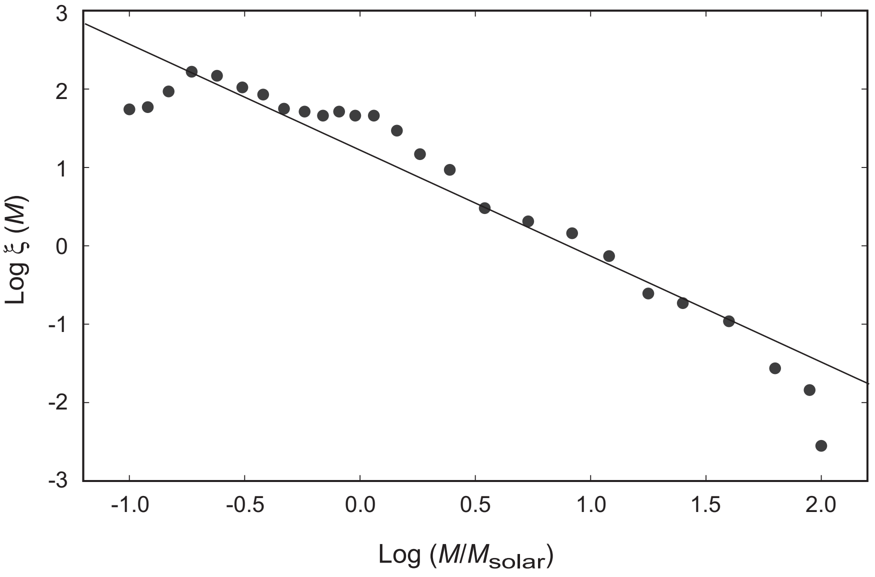


Figure 9.12: Initial mass function (IMF). Points are for stars near the Sun and the line represents a Saltpeter power law, $\log \xi(M) = -1.35 \log M + 1.2$.

Salpeter found a simple power law,

$$\xi(M) = \xi_0 M^{-1.35}.$$

The IMF for stars near the Sun from more recent work is shown in Fig. 9.12, along with Salpeter's estimate.

- Salpeter's power law continues to work well for stellar masses in the $\sim 0.2 - 80 M_\odot$ range.
- Deviation at large M likely is the *mass limit on stars*. At small M there are *complexities not in the Salpeter model*.
- Clearly *massive stars are rare,* and the vast majority of stars are produced with masses well below $1M_\odot$.
- The most likely result of star formation is *a main sequence star with a mass of a few tenths of a $M_\odot$.*

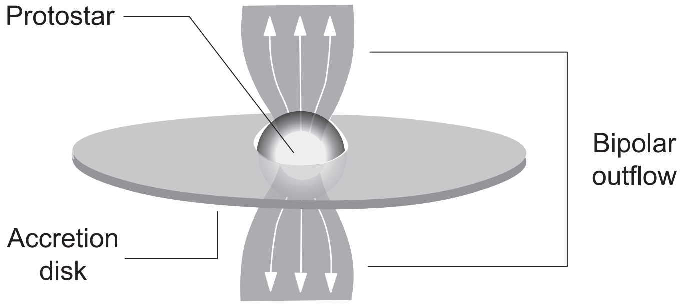


Figure 9.13: Schematic model of an accretion disk and bipolar outflow.

9.13 Protoplanetary Disks

- In the final stages of protostar collapse, matter will continue to accrete from an equatorial accretion disk.
- Young stars produce very strong stellar winds that are focused perpendicular to the equatorial accretion disk.
- Accretion disks and bipolar outflow may common for stars collapsing to the main sequence. (Fig. 9.13).
- The strong wind blowing from young stars is not well understood. One possible cause is matter drawing a magnetic field inward as it falls into the accretion disk.
- The outward flowing wind is partially blocked by the accretion disk and so escapes along the polar axis, producing bipolar outflows from the young star.
- However, this simple picture cannot explain the narrow width of the outflow observed in some cases. Presumably the full mechanism is more complex, perhaps involving the effect of magnetic fields to focus the ejected material.

9.14 Stars, Disks, and Angular Momentum

The preceding discussion of collapsing protostars has mentioned the role of angular momentum only in passing.

- The interstellar clouds from which protostars collapse will in general be rotating slowly.
- Doppler shifts of radio waves from opposite sides of these clouds suggest line-of-sight velocities of order $0.1\,\mathrm{km\,s^{-1}}$.

If such a cloud collapses decoupled from the rest of the Universe *its angular momentum must be preserved* so $v_0 r_0 = v_\mathrm{f} r_\mathrm{f}$,

- where v_0 and r_0 denote an initial tangential velocity and radius, respectively, and
- v_f and r_f denote the corresponding quantities after the collapse.

Earlier it was estimated that a $1\,M_\odot$ cloud collapses to a star from an initial radius of order 10^{16} cm.

- Taking the Sun as representative, this corresponds to a decrease in radius by 5–6 orders of magnitude.
- Invoking conservation of angular momentum, if the $1\,M_\odot$ cloud collapsed directly to the Sun from
  - an initial radius of $10^{16}\,\mathrm{cm}$ and
  - tangential velocity $0.1\,\mathrm{km\,s^{-1}}$,
- the surface of the Sun should have been spun up to a velocity

$$v = \left(\frac{r_0}{R_\odot}\right) v_0 \simeq 14{,}000\,\mathrm{km\,s^{-1}}.$$

- No normal star is spinning at anywhere near this rate!

A basic fallacy in the preceding argument is that for finite angular momentum

- The collapse will not proceed directly to a star.
- Instead it will terminate when the rotating cloud forms a stable disk for which the gravitational acceleration exactly keeps the particles in a circular orbit
- This requires that
$$\frac{v_\mathrm{f}^2}{r_\mathrm{f}} = \frac{GM}{r_\mathrm{f}^2},$$
which may be combined with
$$v = \left(\frac{r_0}{R_\odot}\right) v_0$$
to give a disk radius
$$r_\mathrm{disk} = r_\mathrm{f} \simeq \frac{v_0^2 r_0^2}{GM}.$$

These disk radii typically are of order 100 AU.

Thus, the initial collapse is likely to a rotating disk much larger than a star, and

- the final star is produced by an object that condenses at the center of this disk
  - having *much of the disk's mass* but
  - only a *fraction of its angular momentum*.
- The mechanism by which this takes place is not well understood but involves *transfer of angular momentum outward in the disk*.

Thus, for example, in our Solar System the outer planets like Jupiter carry much more angular momentum in their orbital motion than the Sun carries in its rotation.

9.15 Exoplanets

The dust disks observed around a number of young stars suggests that planetary formation may be taking place in these systems.

- Indeed, over the past few years impressive evidence has accumulated for thousands of *extrasolar planets* or *exoplanets*.
- These planets are difficult to observe directly at their great distance.
- They are detected primarily through their gravitational influence on the parent star.
- In principle they could be detected by the wobble in angular position on the celestial sphere of the parent star caused by the gravitational tug of the planet as it moves about its orbit.
- In practice, it has proven easier to instead
  - detect the wobble corresponding to the small periodic Doppler shifts for the radial motion of the parent star induced by this motion, and (in favorable cases)
  - by small variations in light output caused by planetary transits of the parent star.

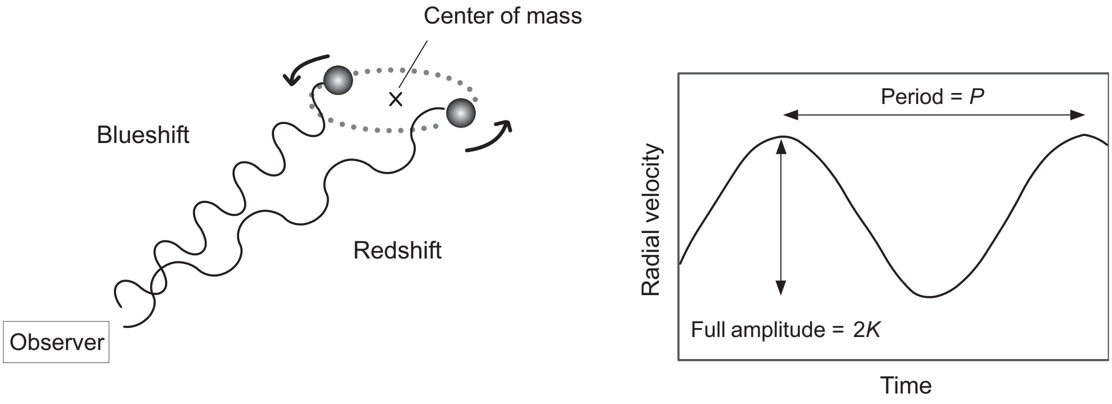


Figure 9.14: Doppler spectroscopy method for detecting extrasolar planets.

9.15.1 The Doppler Spectroscopy Method

The *Doppler spectroscopy method* is illustrated in Fig. 9.14.

- he semiamplitude K of the radial velocity is

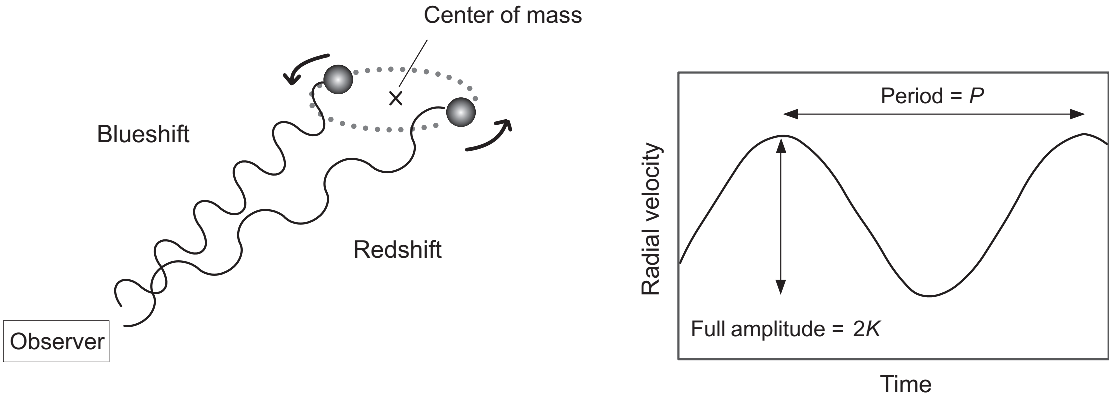
$$K = \left(\frac{2\pi G}{P}\right)^{1/3} \frac{m_{\rm p} \sin i}{(M_* + m_{\rm p})^{2/3}(1-e^2)^{1/2}},$$

  - where i is the *tilt angle*,
  - $m_{\rm p}$ is the *mass of the unseen companion*,
  - M_* is the *mass of the observed star*,
  - the *orbital eccentricity* is e,

  and the *orbital period* P is given by *Kepler's third law*,

$$P^2 = \frac{4\pi^2 a^3}{G(M_* + m_{\rm p})},$$

  where a is the *semimajor axis* of the relative orbit.



- Generally, the tilt angle i is unknown, so masses are uncertain by a factor $\sin i$ in the absence of further information.
- The method requires that changes in radial velocity for the parent star of order $10\,\mathrm{m\,s^{-1}}$ be detected.

9.15.2 Transits of Extrasolar Planets

In cases where the geometry is favorable for an eclipse,

- it is possible to detect the transit of extrasolar planets across the face of their parent star through the periodic reduction in light output for the system, and
- in favorable cases the secondary eclipse of the exoplanet by the parent star can be seen.
- Such data allow the tilt angle i of the orbit to be constrained to near $\frac{\pi}{2}$, and
- from eclipse timing the planetary radius can be estimated.
- The IR flux from the planet can be deduced from the total flux decrease in the secondary eclipse, and
- by fitting such data to models, properties of the planet's atmosphere may be inferred.
- Transit information, coupled with the information from Doppler analysis of the system, allows a rather full picture to be constructed:
  - a detailed orbit of the planet,
  - its mass,
  - its size,
  - its density,
  - information about the atmosphere of the planet.

Chapter 10

Life and Times on the Main Sequence

In the preceding chapter we considered the collapse of a protostar to the main sequence. In this chapter we examine the nature of life on the main sequence for such a star.

- The essence of life on the main sequence is *stable burning of H into He in hydrostatic equilibrium*, by
  - *PP chains* for stars of a solar mass or less, and
  - the *CNO cycle* for more massive main-sequence stars.
- Since we have examined in some detail hydrostatic equilibrium, energy production by the PP chains and the CNO cycle, we already understand the essence of life on the main sequence.

It is appropriate that we begin by examining this main sequence star that we know the best: *the Sun*.

10.1 The Standard Solar Model

The Sun is by far the most studied star. This has allowed the construction of a *Standard Solar Model:*

Standard Solar Model: a mathematical model of the Sun that uses

1. fundamental knowledge from fields such as nuclear and atomic physics,
2. measured key quantities, and
3. a few assumptions

to describe all solar observations.

Standard Solar Models are important because

- they fix
  - the Sun's *helium abundance* and
  - the *convection length scale* in the solar sub-surface.
- They provide a *benchmark* for measuring improved solar modeling and a starting point for more general stellar modeling.

The essence of the Standard Solar Model is that *a 1 $M_\odot$ ZAMS star is evolved to the present age of the Sun* subject to the following assumptions:

1. The Sun was *formed from a homogeneous mixture of gases*.
2. The Sun is *powered by nuclear reactions* in its core.
3. The Sun is *approximately in hydrostatic equilibrium*, with the gravitational forces that attempt to compress it almost exactly compensated by forces arising from gradients in internal gas and radiation pressure.
4. Some deviation from equilibrium is permitted as the Sun evolves, but *any deviations are assumed to be small and slow*.
5. Energy is transported from the core, where it is produced, to the surface, where it is radiated into space
    - by photons (*radiative transport*)
    - by large-scale vertical motion of packets of hot gas (*convection*).
6. Any *heat transport by conduction is ignored*.

Let us now discuss each of these assumptions that enter the Standard Solar Model in a little more depth.

10.1.1 Composition of the Sun

- The assumption that the Sun was formed from a *homogenous mixture of gases* is motivated by the *strong convection expected in the protostar* during collapse to the main sequence.
- The surface abundances are then assumed to have been undisturbed in the subsequent evolution, so that

> *Present surface abundances indicate the composition of the original solar core.*

- *The abundance of most elements in the surface can be inferred by spectroscopy.* Exceptions are the noble gases He, Ne, and Ar. They are not excited significantly by the blackbody emission spectrum of the photosphere, so their abundance cannot be determined well from the spectrum.
- Because evolution of the Sun's luminosity depends on the mean molecular weight raised to the power 7.5 (see Exercise 10.1), which is strongly influenced by the helium abundance, *the H/He ratio is normally taken as an adjustable parameter* in solar models.
- The *H/He ratio* is determined by requiring that the luminosity of the Sun at the present age of the Solar System (4.6 billion years, as determined by dating meteorites) be accurately reproduced by the model.

10.1.2 Nuclear Energy Generation and Composition Changes

The Sun is assumed to derive its power and associated composition changes from the proton–proton chains *PP-1*, *PP-2*, and *PP-III*, and the *CNO cycle*. The nuclear reaction networks describing this energy and element production are solved by

- dividing the Sun into concentric shells,
- calculating the nuclear reactions in each shell as a function of the current temperature and density there, and
- using the updated composition and the energy production as constant input to the partial differential equations describing the solar equilibrium structure.

10.1.3 Hydrostatic Equilibrium

Since the dynamical timescale of the Sun is less than an hour,

$$\tau_{\text{hydro}} \simeq (G\bar{\rho})^{1/2} \simeq 55 \text{ minutes},$$

the Sun may be expected to have reached hydrostatic equilibrium quickly.

- However, a Standard Solar Model allows *small expansions and contractions* in response to time evolution of the star.
- *Re-equilibration is assumed to be very fast* compared with the timescale for evolution.
- The pressure is
  - composed of both *gas pressure* and *radiation pressure*, but
  - the radiation pressure even at the center is only about 0.05% of the total pressure.
- A Standard Solar Model typically ignores the effects of
  - rotation,
  - magnetic-field pressure, and
  - stellar pulsations

  on hydrostatic equilibrium.

10.1.4 Energy Transport

It is assumed that (1) energy transport in the Sun by acoustic or gravitational waves is negligible, and that (2) the energy produced internally in the Sun is transported by radiative diffusion and convection to the surface.

- In the interior, transport is assumed to be by
  - *radiative diffusion* unless the critical gradient for convective instability is exceeded, in which case
  - the Sun transports energy in that region *convectively with an adiabatic temperature gradient.*
- In the sub-surface, the actual gradient is steeper than the adiabatic gradient and the resulting convection is modeled by *mixing length theory*.
- Because convection in the sub-surface is difficult to calculate reliably, the mixing length in units of the scale height is taken as an *adjustable parameter,* fixed by requiring the model to yield the observed radius of the Sun.
- The opacities required for radiative diffusion of energy are *Rosseland mean opacities*, calculated numerically.

> Opacities are among the least well-determined quantities entering the Standard Solar Model, with typical uncertainties in the 10–20% range.

Optical Depth and the Solar Surface

The *optical depth* τ at radius r is defined in terms of the radiative opacity κ by

$$\tau = \int_r^\infty \kappa\rho\, dr.$$

It measures the *probability that photons interact with solar material before being radiated into space*.

- The *radius of the Sun* is defined to be that distance from the center where *the optical depth is* $\frac{2}{3}$.
- The diffusion approximation for radiative transport fails when τ is lower than about 1–10 because the mean free path for photons then becomes very long
- (In the solar surface, the mean free path for photons is of order 10^7 cm or longer, compared with fractions of a cm in the interior).
- The region of the solar surface where optical depth is less than about 1 is called the *solar atmosphere.*
- Methods used to deal with radiative transport in the solar atmosphere are much more complicated that those adequate for the solar interior because one can no longer make a diffusion approximation.
- It is *essential to model the atmosphere adequately* because
  - it defines *outer boundary conditions* and
  - this is where the *solar spectrum* is produced.

10.1.5 Constraints and Solution

Solution of the Standard Solar Model problem corresponds to

- evolving in time four partial differential equations in five unknowns (P, T, r, $m(r)$, and L),
- supplemented by an *equation of state* and equations governing *composition change* (one for each species),
- subject to constraints that calculated *radius, luminosity, and mass* are equal to the corresponding observed values.

Two equations governing hydrostatic equilibrium,

$$\frac{dm}{dr} = 4\pi r^2 \rho(r) \qquad \text{Mass conservation}$$

$$\frac{dP}{dr} = -\frac{Gm(r)}{r^2}\rho \qquad \text{Hydrostatic equilibrium,}$$

three equations for luminosity and temperature gradients,

$$\frac{dL}{dr} = 4\pi r^2 \varepsilon(r) \qquad \text{Luminosity}$$

$$\frac{dT}{dr} = -\frac{3\rho(r)\kappa(r)}{4acT^3(r)}\frac{L(r)}{4\pi r^2} \qquad \text{(If radiative)}$$

$$\frac{dT}{dr} = \left(\frac{\gamma-1}{\gamma}\right)\frac{T}{P}\frac{dP}{dr} \qquad \text{(If convective),}$$

equations governing composition changes,

$$\frac{dn}{dt} = -\frac{1}{\tau}n \qquad \text{Nucleosynthesis,}$$

and an equation of state,

$$P = P(T, \rho, X_i, \ldots) \qquad \text{Equation of state.}$$

The network of equations required to describe nuclear energy and element production is solved separately for each timestep in each zone.

- The equation of state is
  - assumed to be given by the *ideal gas law for regions that are completely ionized*.
  - Otherwise, a *numerical equation of state* is typically used.
- The Standard Solar Model solution is constructed iteratively.
  - Starting values for the helium abundance and the mixing length parameter are used to evolve an initial zero-age model to the current age of the Sun.
  - The model's luminosity and radius are then compared with observations, the helium abundance and mixing length parameters adjusted accordingly, and the model is evolved again.
  - This cycle is repeated until convergence is obtained.

Table 10.1 gives the temperature, density, pressure, and luminosity of a Standard Solar Model as a function of radius and enclosed mass at that radius.

Table 10.1: A Standard Solar Model

$M/M_\odot$	$R/R_\odot$	T (K)	ρ (g cm^{-3})	P (dyn cm^{-2})	$L/L_\odot$
0.0000298	0.00650	1.568E+07	1.524E+02	2.336E+17	0.00027
0.0008590	0.02005	1.556E+07	1.483E+02	2.280E+17	0.00753
0.0065163	0.04010	1.516E+07	1.359E+02	2.111E+17	0.05389
0.0207399	0.06061	1.456E+07	1.193E+02	1.868E+17	0.15638
0.0439908	0.08041	1.386E+07	1.027E+02	1.606E+17	0.29634
0.0762478	0.10006	1.310E+07	8.729E+01	1.349E+17	0.45135
0.1173929	0.12000	1.231E+07	7.350E+01	1.108E+17	0.60142
0.1672004	0.14056	1.150E+07	6.123E+01	8.892E+16	0.73152
0.2203236	0.16027	1.076E+07	5.114E+01	7.094E+16	0.82657
0.2800107	0.18104	1.002E+07	4.205E+01	5.517E+16	0.89658
0.3393826	0.20107	9.353E+06	3.459E+01	4.279E+16	0.94011
0.3966733	0.22038	8.762E+06	2.847E+01	3.319E+16	0.96616
0.4559683	0.24084	8.188E+06	2.301E+01	2.516E+16	0.98259
0.5114049	0.26085	7.676E+06	1.857E+01	1.907E+16	0.99183
0.5627338	0.28058	7.214E+06	1.496E+01	1.446E+16	0.99669
0.6099028	0.30016	6.794E+06	1.203E+01	1.096E+16	0.99860
0.6564038	0.32132	6.379E+06	9.484E+00	8.119E+15	0.99941
0.6952616	0.34091	6.028E+06	7.605E+00	6.156E+15	0.99976
0.7304369	0.36063	5.703E+06	6.092E+00	4.667E+15	0.99993
0.7621708	0.38053	5.400E+06	4.876E+00	3.539E+15	1.00002
0.7907148	0.40067	5.117E+06	3.900E+00	2.683E+15	1.00005
0.8163208	0.42109	4.851E+06	3.118E+00	2.034E+15	1.00007
0.8374222	0.44008	4.621E+06	2.539E+00	1.578E+15	1.00007
0.8580756	0.46112	4.383E+06	2.029E+00	1.197E+15	1.00006
0.8750244	0.48072	4.176E+06	1.651E+00	9.287E+14	1.00006
0.8902432	0.50063	3.978E+06	1.345E+00	7.206E+14	1.00005
0.9038831	0.52086	3.789E+06	1.095E+00	5.591E+14	1.00004
0.9160850	0.54139	3.606E+06	8.924E-01	4.339E+14	1.00004
0.9260393	0.56033	3.445E+06	7.413E-01	3.445E+14	1.00003
0.9358483	0.58142	3.273E+06	6.052E-01	2.673E+14	1.00003

Table 10.1: (Continued) Standard Solar Model

$M/M_\odot$	$R/R_\odot$	$T\,$(K)	$\rho\,$(g cm^{-3})	$P\,$(dyn cm^{-2})	$L/L_\odot$
0.9438189	0.60081	3.120E+06	5.040E-01	2.123E+14	1.00002
0.9509668	0.62036	2.969E+06	4.205E-01	1.686E+14	1.00002
0.9573622	0.64001	2.818E+06	3.517E-01	1.339E+14	1.00002
0.9636045	0.66168	2.648E+06	2.900E-01	1.039E+14	1.00001
0.9686223	0.68129	2.485E+06	2.445E-01	8.249E+13	1.00001
0.9730081	0.70042	2.315E+06	2.081E-01	6.572E+13	1.00001
0.9771199	0.72033	2.115E+06	1.780E-01	5.161E+13	1.00001
0.9811002	0.74162	1.899E+06	1.513E-01	3.936E+13	1.00000
0.9842836	0.76050	1.718E+06	1.299E-01	3.055E+13	1.00000
0.9874435	0.78148	1.526E+06	1.085E-01	2.264E+13	1.00000
0.9900343	0.80103	1.355E+06	9.066E-02	1.678E+13	1.00000
0.9922832	0.82051	1.193E+06	7.470E-02	1.215E+13	1.00000
0.9942853	0.84082	1.031E+06	5.987E-02	8.406E+12	1.00000
0.9958822	0.86022	8.826E+05	4.733E-02	5.682E+12	1.00000
0.9972278	0.88035	7.356E+05	3.590E-02	3.585E+12	1.00000
0.9982619	0.90020	5.966E+05	2.613E-02	2.110E+12	1.00000
0.9990296	0.92017	4.627E+05	1.775E-02	1.107E+12	1.00000
0.9995498	0.94015	3.343E+05	1.080E-02	4.833E+11	1.00000

$M_\odot = 1.989 \times 10^{33}$ g $R_\odot = 6.96 \times 10^{10}$ cm $L_\odot = 3.827 \times 10^{33}$ erg s^{-1}

Figure 10.1 illustrates graphically some of the parameters of this model plotted versus the radius and Fig. 10.2 plots the same quantities versus the enclosed mass coordinate.

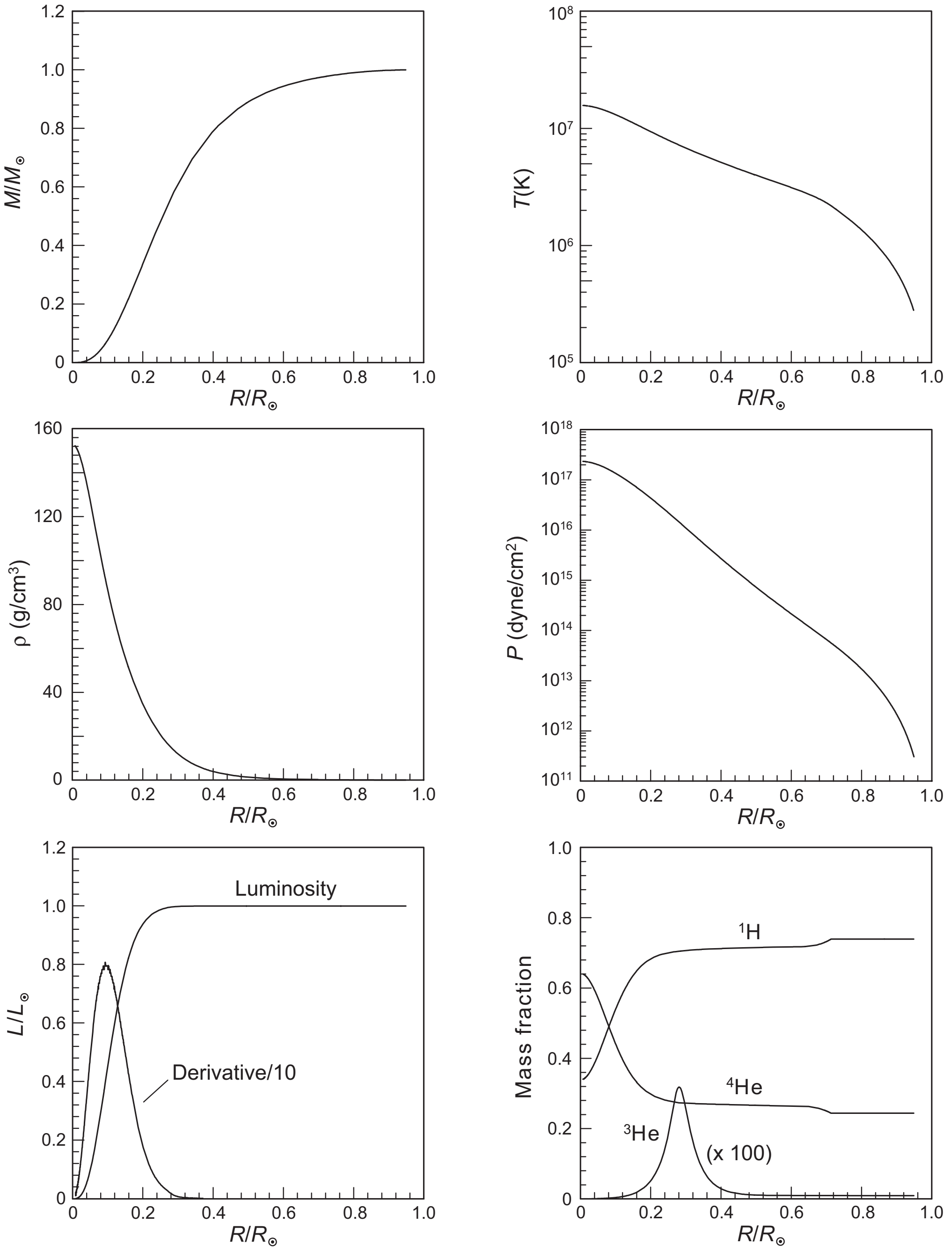


Figure 10.1: Parameters from a Standard Solar Model (Table 10.1) plotted versus the radial coordinate.

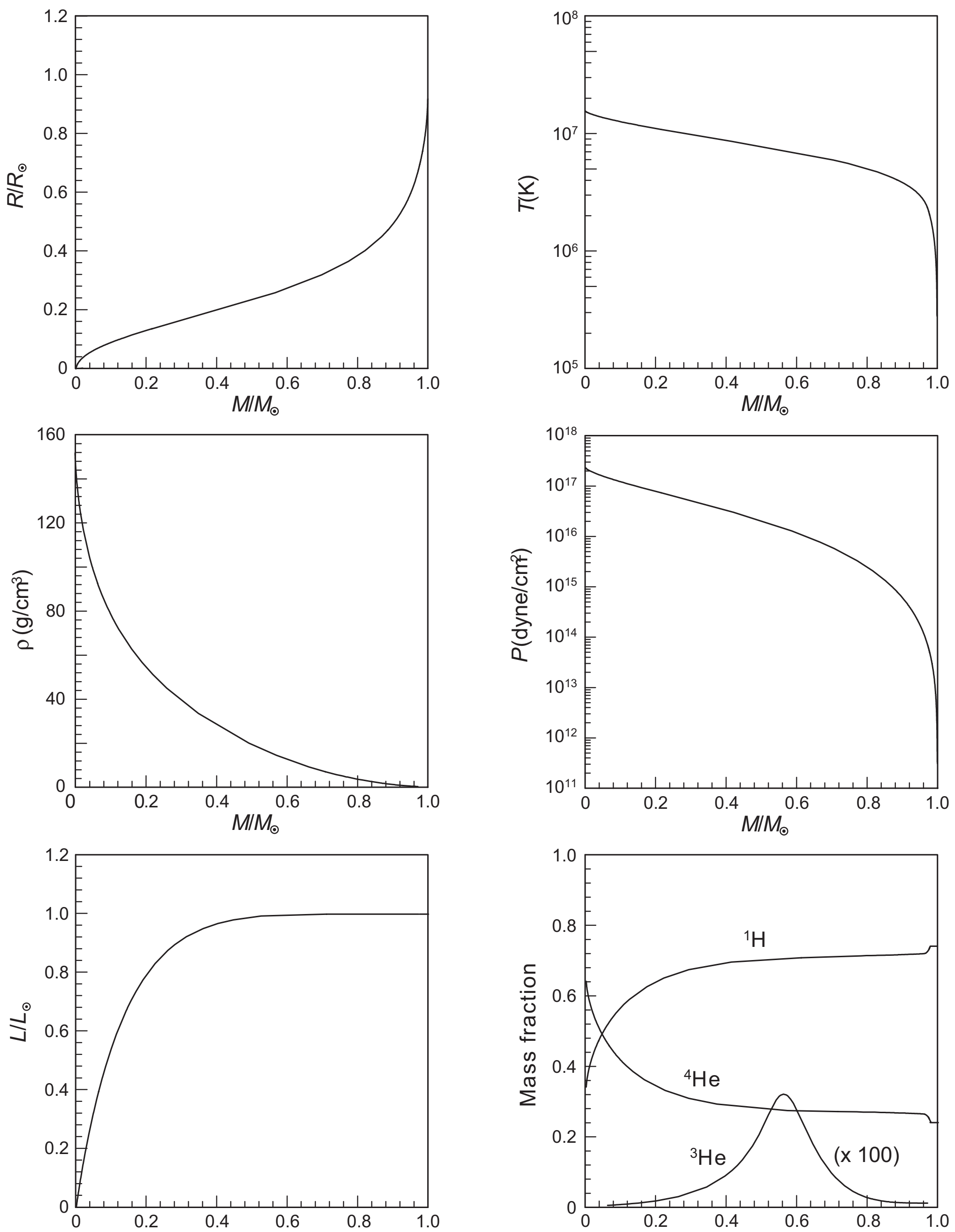


Figure 10.2: Parameters from a Standard Solar Model (Table 10.1) plotted versus the enclosed mass.

The Standard Solar Model may be tested by comparing its predictions with observations.

- These tests range from general ones, such as accounting for the existence, age, and energy output of the Sun, to specific ones such as the accounting for the results of solar seismology.
- The Standard Solar Model has passed these tests very well.

We now discuss two examples of how the Standard Solar Model description of the solar interior can be tested:

- The subsurface structure as inferred from helioseismology and
- The spectrum and overall flux of neutrinos emitted from the solar core.

10.2 Helioseismology

One way to study the Sun's interior is to study the propagation of waves in its body.

- This is similar to the way geologists learn about the interior of the Earth by studying seismic waves or how we may infer the composition of a bell by striking it and studying the sound frequencies that it produces.
- The corresponding field of study is called *helioseismology.*

10.2.1 p-Modes

Solar oscillations were discovered by studying Doppler shifts of surface absorption lines.

- It was found that the solar surface consists of patches oscillating on a timescale of five minutes with a velocity amplitude of $0.5\,\mathrm{km\,s^{-1}}$.
- These 5-minute oscillations represent pressure waves (*p-modes*) trapped between the surface and the lower boundary of the convective zone.
- They are *reflected* from the solar surface by density gradients.
- They are *refracted* near the bottom of the convection zone because of changing soundspeed in that region.

10.2.2 g-Modes

In addition to p-modes associated with acoustical waves trapped near the solar surface, the Sun may also exhibit *g-modes:*

- These correspond to oscillations in which the restoring force is gravity.
- If g-modes can be observed, they carry information about much deeper regions of the Sun than that carried by the p-modes.

The Sun vibrates at a complex set of frequencies, with the dominant frequency corresponding to the 5-minute oscillation described above.

- By decomposing the observed vibration of the Sun into a superposition of standing acoustic waves, it is possible to learn about the interior.
- Such decompositions indicate that the observed motion of the surface is a superposition of several million resonant modes with different frequencies and horizontal wavelengths.
- Individual modes in this decomposition may have velocity amplitudes as large as $20\,\mathrm{cm\,s^{-1}}$ and 1–2 meter vertical displacements.

Presently, helioseismology is placing strong constraints on our theories of the solar interior.

- The analysis is complex but the basic idea is simple: *changes in the properties of the solar interior (for example, the amount of helium in some region) affects the way sound waves travel through the interior.*
- This will in turn influence the way the surface vibrates.

Two important pieces of information obtained from early helioseismology are that

- the abundance of helium in the interior (but outside the core) is about the same as at the surface, and that
- convection extends about 30 percent of the way to the center.

However, more recently it has been found that stellar evolution models applied to the Sun

- are compatible with helioseismology if they adopt older solar composition models, but
- are *incompatible with helioseismology* if they adopt the solar composition obtained by the *newest spectroscopic models*.

This unresolved anomaly is called the *solar abundance problem*.

10.3 Solar Neutrino Production

Helioseismology allows us to probe the interior of the Sun. A second way in which we can study the (deep) interior of the Sun is by detecting the neutrinos that are produced there.

- The energy powering the surface photon luminosity must make its way on a *100,000-year or greater timescale* to the surface before being radiated.
- Neutrinos emitted from the core are largely unimpeded in their exit from the Sun, reaching the Earth 8.5 minutes after they were produced.

Therefore, neutrinos carry *immediate and more direct information* about the current conditions in the solar core than do the photons emitted from the solar photosphere.

Eight reactions or decays of some significance in solar energy production produce neutrinos:

pp	$p + p \rightarrow {}^2H + e^+ + \nu_e$	$Q \leq 0.420\,\text{MeV}$
pep	$p + e^- + p \rightarrow {}^2H + \nu_e$	$Q = 1.442\,\text{MeV}$
hep	${}^3He + p \rightarrow {}^4He + \nu_e$	$Q \leq 18.773\,\text{MeV}$
^{7}Be	${}^7Be + e^- \rightarrow {}^7Li + \nu_e$	$Q = 0.862\,\text{MeV}\ (89.7\%)$
		$Q = 0.384\,\text{MeV}\ (10.3\%)$
^{8}B	${}^8B \rightarrow {}^8Be^* + e^+ + \nu_e$	$Q \leq 15\,\text{MeV}$
CNO	${}^{13}N \rightarrow {}^{13}C + e^+ + \nu_e$	$Q \leq 1.199\,\text{MeV}$
CNO	${}^{15}O \rightarrow {}^{15}N + e^+ + \nu_e$	$Q \leq 1.732\,\text{MeV}$
CNO	${}^{17}F \rightarrow {}^{17}O + e^+ + \nu_e$	$Q \leq 1.740\,\text{MeV}$

- Six of the reactions produce spectra with a range of Q-values.
- Two are line sources (produce sharp energies).
- CNO neutrinos are difficult to detect because
  - *intensities are weak* (less than 2% of Sun's energy)
  - the *energies are low*.
- Therefore, our primary concern will be with the *first five reactions*, which correspond to *steps of the PP chains*.

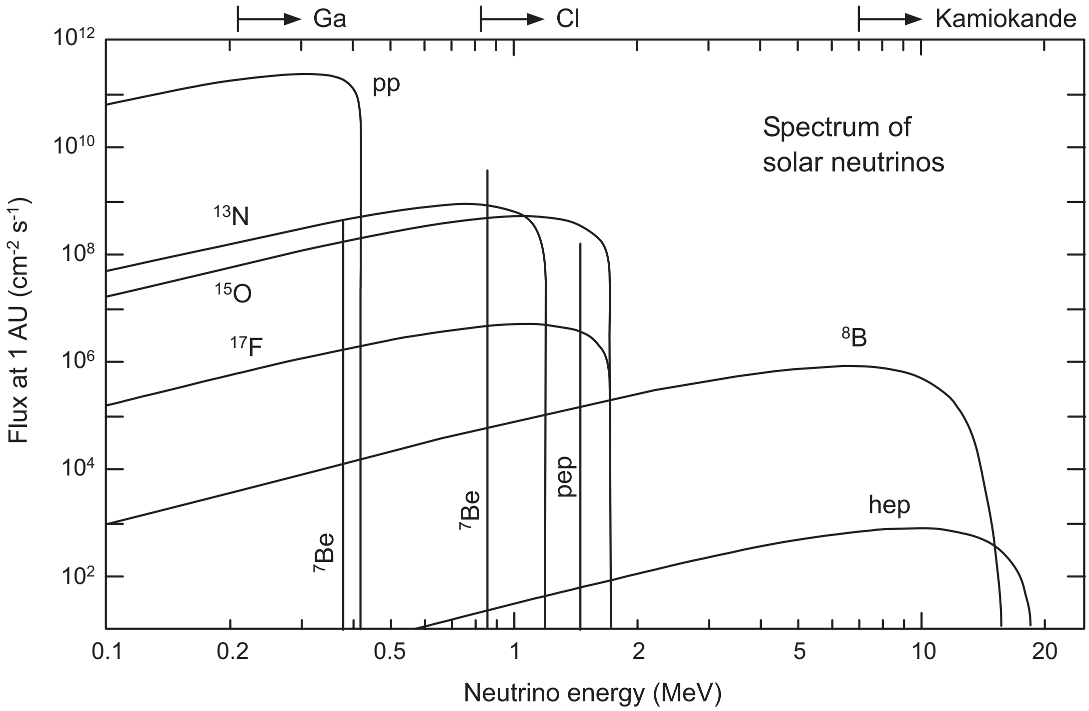


Figure 10.3: The solar neutrino spectrum. The sensitive region of various experiments is indicated above the graph.

The solar neutrino spectrum predicted by the Standard Solar Model is shown in Fig. 10.3.

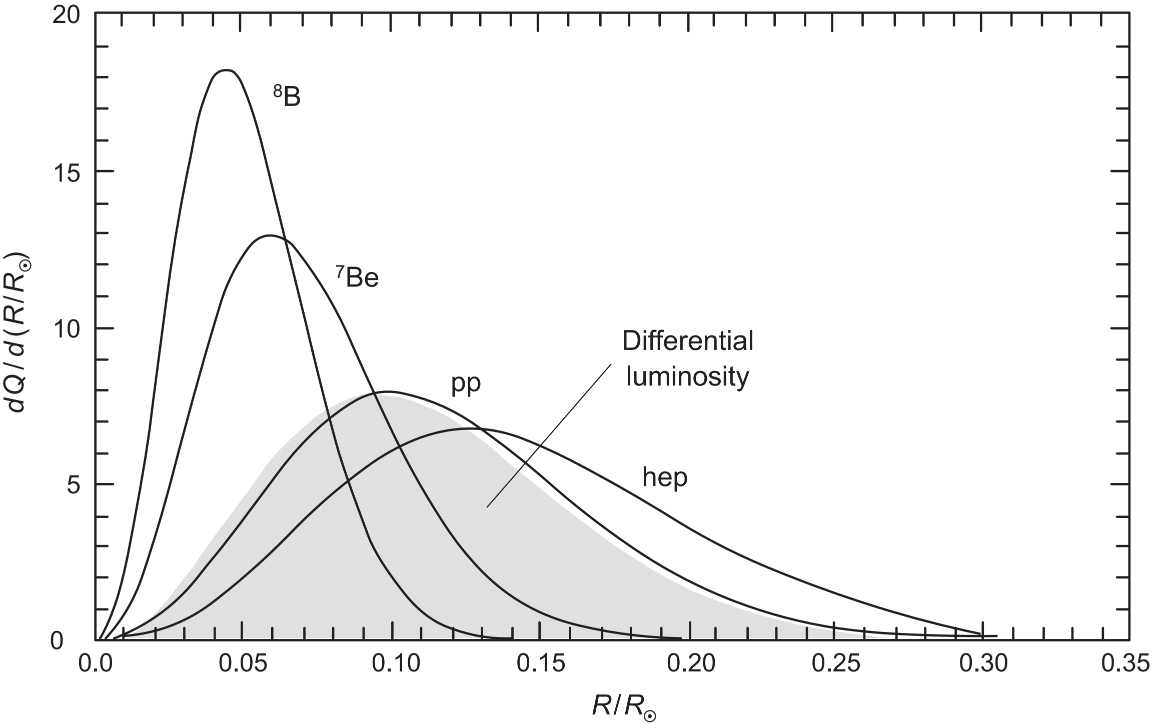
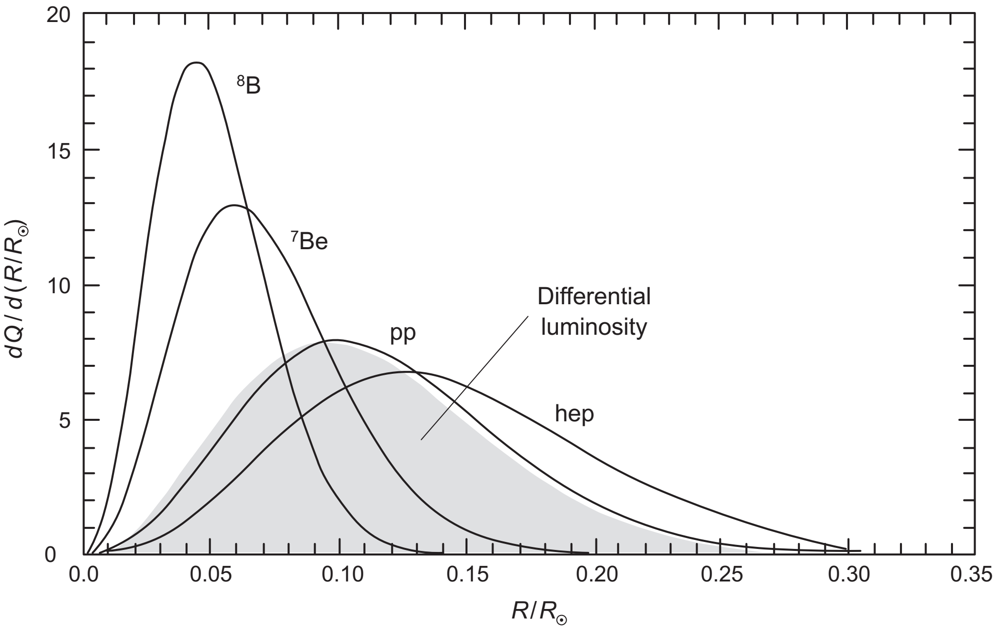


Figure 10.4: Differential neutrino production dQ/dR as a function of solar radius. The shaded area indicates the differential photon luminosity.

Fig. 10.4 illustrates the radial regions of the Sun responsible for producing neutrinos from each of the PP reactions.



- The ^{8}*B and* ^{7}*Be neutrinos probe smaller radii* than the neutrinos produced in PP-I (labeled pp).
- Since they are *produced at smaller* ***R*** *they are produced at higher* ***T***.
- Attempts to understand the rate of observed neutrino emission from the Sun yielded *initially surprising results*.
- These results suggested that our fundamental understanding of either (or both)
  - *elementary particle physics* or
  - *how the Sun works*

  were incomplete.
- Resolution of this issue led to profound new understanding in astrophysics and elementary particle physics.

10.4 The Solar Electron-Neutrino Deficit

By counting the number of neutrinos produced and the average energy released for each $4\text{H} \rightarrow {}^4\text{He}$ in the PP chains, we may estimate that

- The Sun should be emitting approximately 10^{38} *electron neutrinos per second* if it is powered by the PP chains.
- However, detectors on Earth see only a fraction of the corresponding number of electron neutrinos that should reach Earth.
- This has historically been termed the *solar neutrino problem.*

Let's now describe the *solar neutrino detection experiments* that led to this conclusion.

10.4.1 The Davis Chlorine Experiment

The pioneering solar neutrino detection experiment was started by *Raymond Davis* in the early 1960s.

- It uses the reaction

$$\nu + {}^{37}\mathrm{Cl} \rightarrow {}^{37}\mathrm{Ar} + e^-$$

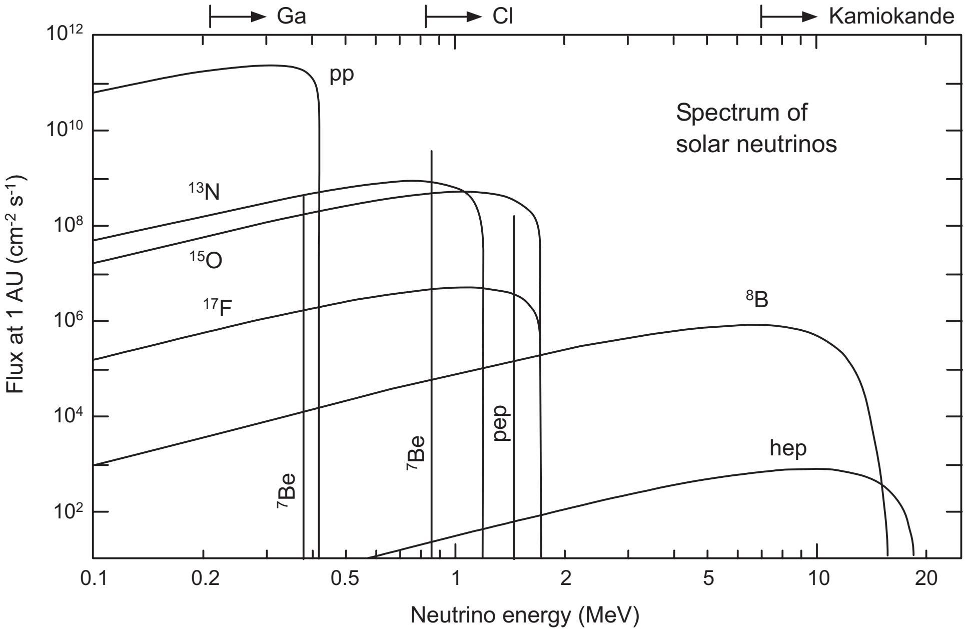
  initiated in *600 tons of cleaning fluid (C_2Cl_4)*.

- To *shield against cosmic rays*, the tank containing the cleaning fluid was placed *1500 meters below the surface* in the abandoned Homestake gold mine in South Dakota.

- The *small number of argon atoms* produced by the above reaction are *radioactive*.

- Their *radioactive decays can be counted* after *separation of the argon from the cleaning fluid by chemical means*.

- The reaction $\nu + {}^{37}\text{Cl} \rightarrow {}^{37}\text{Ar} + e^-$ has a threshold (minimum energy for the reaction to occur) of 0.8 MeV.
- This is higher than the maximum energy of 0.42 MeV for neutrinos in the PP-I chain:



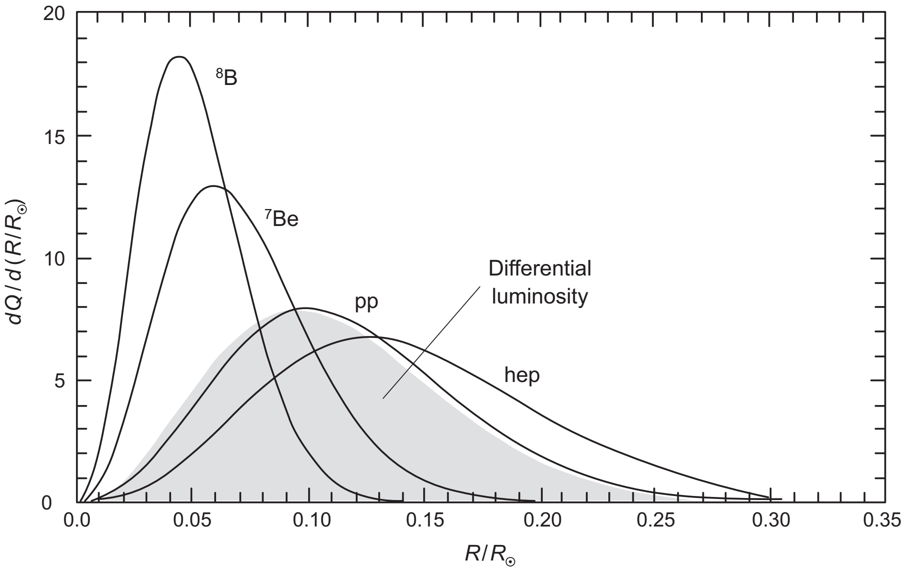
- Therefore, *the Davis experiment was sensitive primarily to the* ^{8}*B neutrinos* (and weakly to the ^{7}Be neutrinos).

> The Davis rate was *2–3 times smaller* than the rate predicted by the Standard Solar Model.

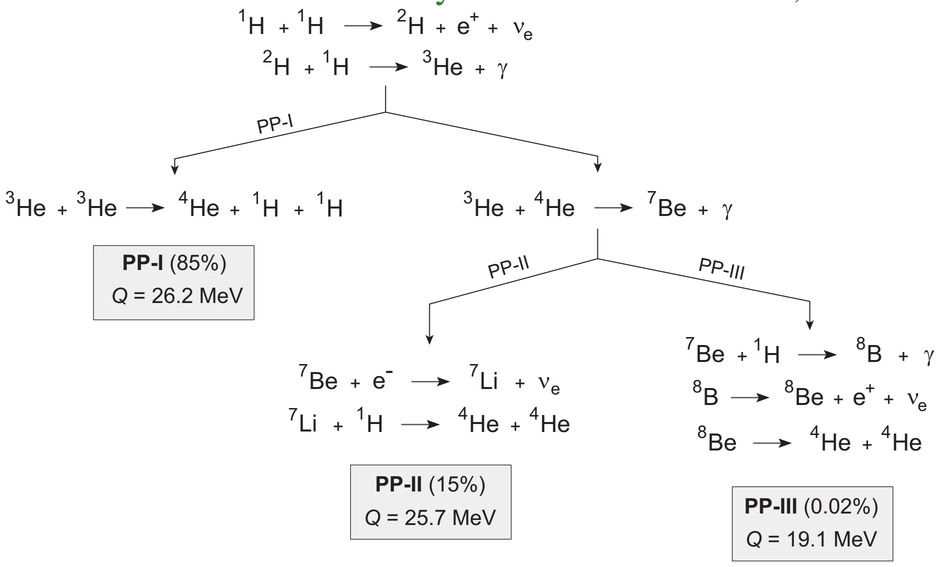
- This detection method had *no directional sensitivity*.

10.4.2 The Gallium Experiments

The Davis experiment was mostly sensitive to the ^{8}B neutrinos.



These come come from a very minor PP side branch,



$$^1\text{H} + {}^1\text{H} \longrightarrow {}^2\text{H} + e^+ + \nu_e$$
$$^2\text{H} + {}^1\text{H} \longrightarrow {}^3\text{He} + \gamma$$

PP-I

$$^3\text{He} + {}^3\text{He} \longrightarrow {}^4\text{He} + {}^1\text{H} + {}^1\text{H}$$

PP-I (85%)
Q = 26.2 MeV

$$^3\text{He} + {}^4\text{He} \longrightarrow {}^7\text{Be} + \gamma$$

PP-II

$$^7\text{Be} + e^- \longrightarrow {}^7\text{Li} + \nu_e$$
$$^7\text{Li} + {}^1\text{H} \longrightarrow {}^4\text{He} + {}^4\text{He}$$

PP-II (15%)
Q = 25.7 MeV

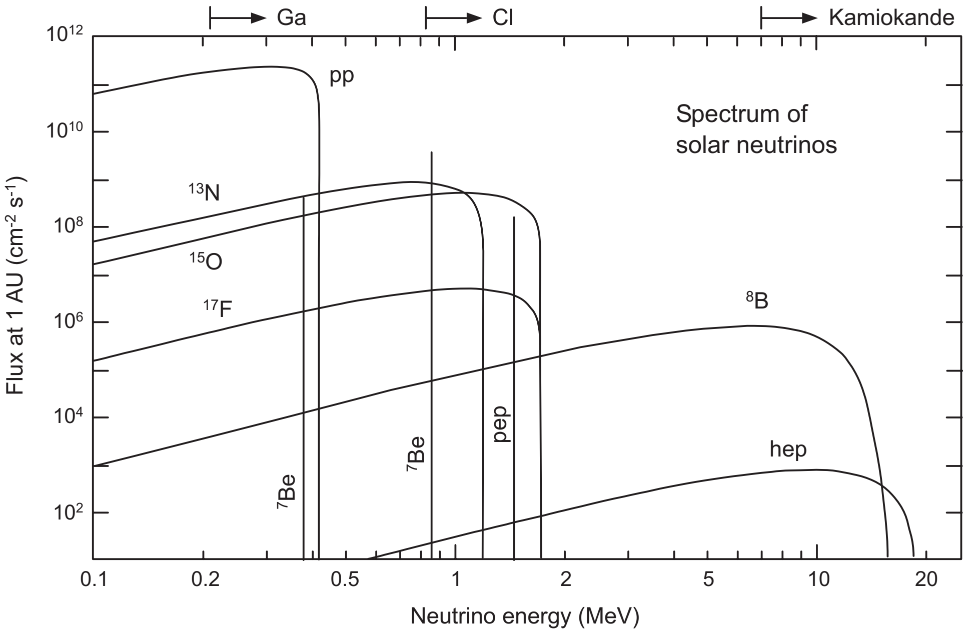
PP-III

$$^7\text{Be} + {}^1\text{H} \longrightarrow {}^8\text{B} + \gamma$$
$$^8\text{B} \longrightarrow {}^8\text{Be} + e^+ + \nu_e$$
$$^8\text{Be} \longrightarrow {}^4\text{He} + {}^4\text{He}$$

PP-III (0.02%)
Q = 19.1 MeV

which is not very closely related to the Sun’s photon luminosity.

Because of the threshold for $\nu + {}^{37}\mathrm{Cl} \rightarrow {}^{37}\mathrm{Ar} + e^-$, neutrinos from PP-I (primary solar energy source) are *not detected at all*.



Another *chemistry-based detector* can be constructed using

$$\nu + {}^{71}\mathrm{Ga} \rightarrow {}^{71}\mathrm{Ge} + e^-.$$

The ^{71}Ge produced is *radioactive*.

- Thus the *Ge can be separated chemically from the Ga* and the decay of ^{71}Ge *counts the number of neutrino reactions*.
- The reaction $\nu + {}^{71}\mathrm{Ga} \rightarrow {}^{71}\mathrm{Ge} + e^-$ has a threshold of only 0.23 MeV.
- Thus it can *detect neutrinos coming from the PP-I chain* that produces most of the solar energy.

Two large experiments,

- *SAGE* (operated by a Russian–American collaboration underground in the Caucasus) and
- *GALLEX* (operated by a largely European collaboration in the Gran Sasso underground laboratory in Italy),

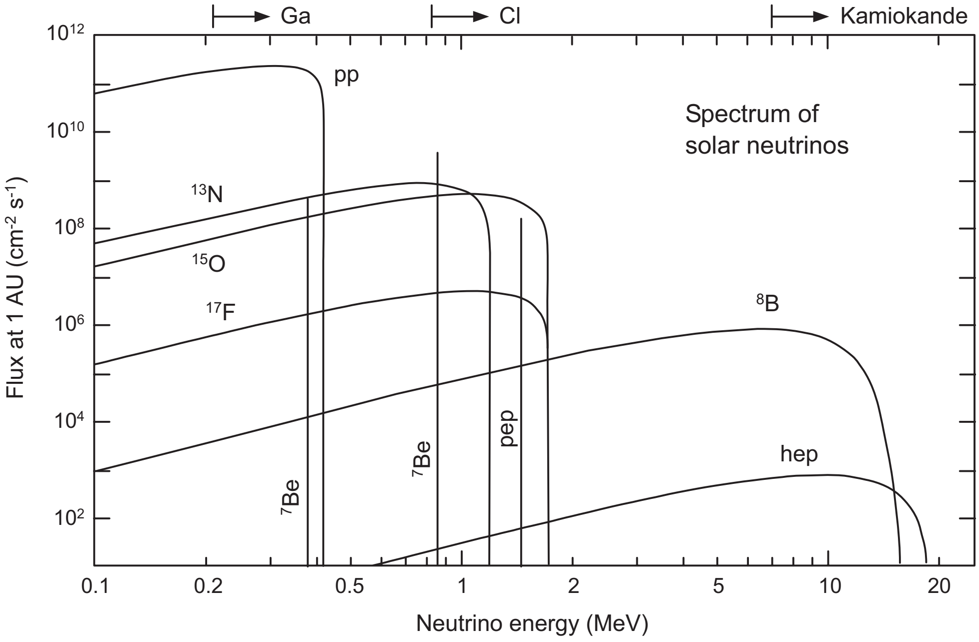
were implemented based on the gallium reaction $\nu + {}^{71}\text{Ga} \to {}^{71}\text{Ge} + e^-$.

- For these experiments *more than half of the neutrinos came from the pp reaction in PP-I*.
- These experiments *also measured a neutrino deficit* compared with the Standard Solar Model.
  - However, the deficit is *not as large as in the Davis chlorine experiment*.
  - They found that the electron neutrino flux is *reduced by a factor of about two* relative to that expected from the Standard Solar Model (SSM).
- Like the chlorine experiment, the gallium experiments were chemistry-based and had *no directional sensitivity*.

10.4.3 Super Kamiokande

The neutrino detector *Super Kamiokande* (commonly referred to as *Super-K*) uses a *different approach.*

- A large tank containing *50,000 cubic meters of ultrapure water* is monitored by photodetectors.
- When the elastic scattering reaction $\nu + e^- \rightarrow \nu + e^-$ occurs in the water, recoiling electrons
- may exceed the speed of light in the medium and produce *Čerenkov radiation* that is detected by the phototubes.
- This detector is sensitive only to the *more energetic* ^8B *neutrinos produced in PP-III.*



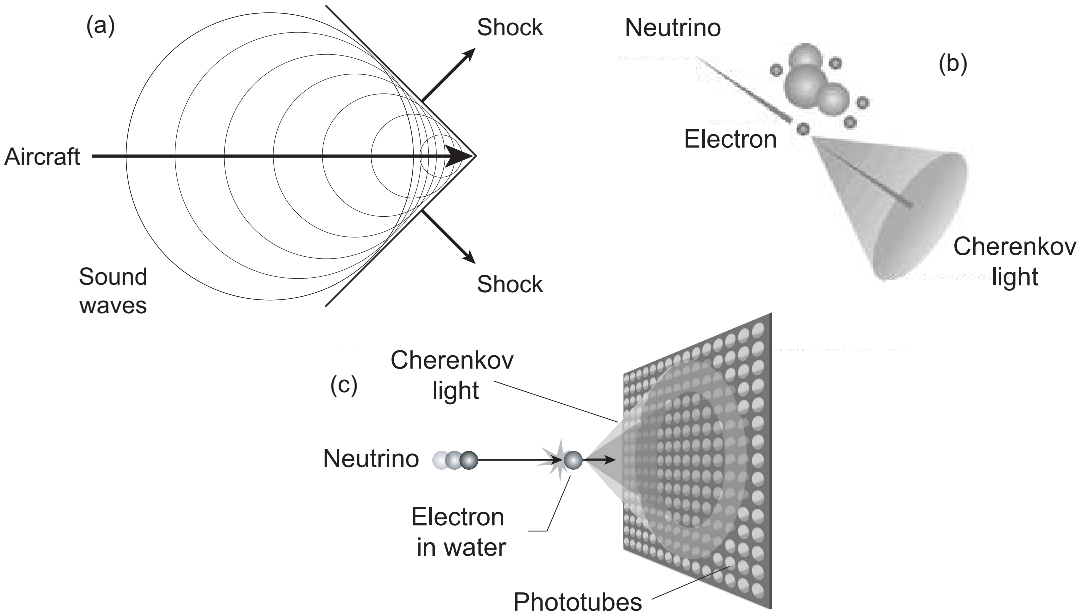


Figure 10.5: (a) Shockwaves produced by exceeding the speed of sound in a medium. (b) Production of Cerenkov radiation by neutrinos. (c) Detection of Cerenkov radiation.

- Super-K has directional sensitivity.
- Thus it is able to demonstrate that the neutrinos being detected *come from the direction of the Sun*.

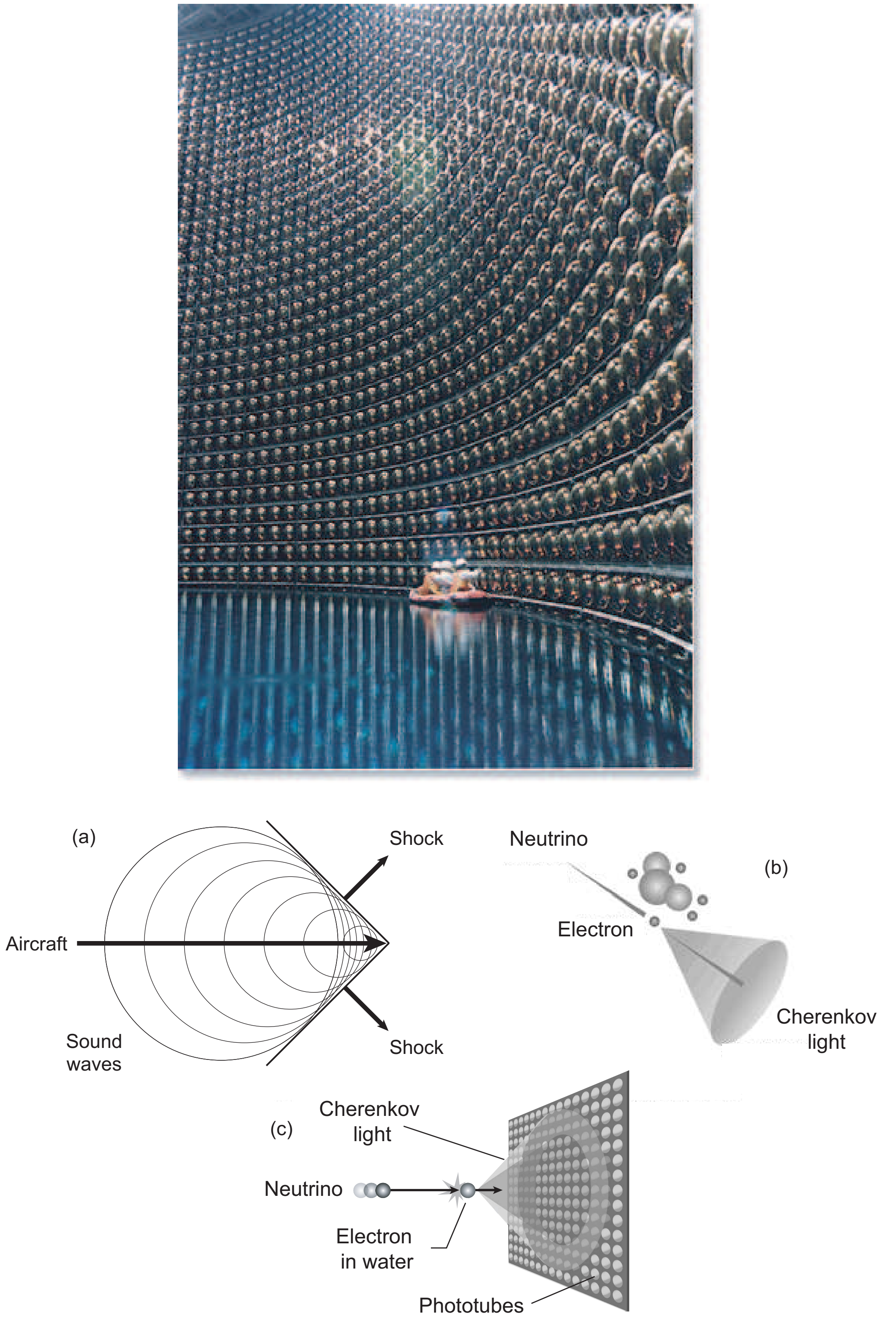


Figure 10.6: (a) (top) Workers inspecting Super-K with the tank partially filled. The bulbs are the photomultiplier tubes. (bottom) Schematic of detecting neutrinos.

- The Super Kamiokande results again *indicated a solar neutrino deficit*:
- The detector sees *fewer than 40% of the electron neutrinos expected* based on fluxes predicted by the Standard Solar Model.

Table 10.2: Solar neutrino fluxes from various experiments compared with a Standard Solar Model (SSM). All fluxes are in solar neutrino units (SNU), except the result from Super Kamiokande. Experimental uncertainties include systematic and statistical contributions.

Experiment	Observed flux	SSM	Observed/SSM
Homestake	$2.54 \pm 0.14 \pm 0.14$ SNU	$9.3\,^{+1.2}_{-1.4}$	0.273 ± 0.021
SAGE	$72\,^{+12}_{-10}\,^{+5}_{-7}$ SNU	$137\,^{+8}_{-7}$	0.526 ± 0.089
GALLEX	$69.7 \pm 6.7\,^{+3.9}_{-4.5}$ SNU	$137\,^{+8}_{-7}$	0.509 ± 0.089
Super-Kamiokande	$2.51\,^{+0.14}_{-0.13}$ $(10^6\,\mathrm{cm}^{-2}\mathrm{s}^{-1})$	$6.62\,^{+0.93}_{-1.12}$	0.379 ± 0.034

The experiments described above were not all sensitive to the same neutrinos from the Sun and found somewhat different magnitudes for the solar neutrino deficit.

- However, the chlorine experiment, two gallium experiments, and water Čerenkov detectors all find reproducibly that *significantly fewer neutrinos are being detected coming from the Sun than the Standard Solar Model predicts*.
- Table 10.2 summarizes and compares with the predictions of a Standard Solar Model (SSM).
- These results indicate
  - a deficit of solar neutrinos in the detectors, and that
  - the deficit depends on which neutrinos are detected.

Example: suppression of ^{8}B neutrinos appears to be larger than for the PP-I neutrinos.

10.4.4 Astrophysics Versus Particle Physics Explanations

Confirmation in more recent experiments of the solar neutrino problem discovered by Davis implies two alternatives:

- We do not understand how the Sun works *(failure of the Standard Solar Model).*
- We do not understand the neutrino *(failure of the Standard Model of elementary particle physics).*

Thus, a debate ensued over whether the solution to the solar neutrino problem lay in a modification of our astrophysics understanding or of our particle physics understanding.

- Experiments and observations have shown rather conclusively that the "solar neutrino problem" is now resolved, and that
- *the resolution lies in new properties for neutrinos* that imply physics beyond the Standard Model of elementary particle physics.
- Specifically, we now have strong evidence that
  - *at least some neutrinos have a non-zero mass,* and
  - this permits neutrinos to change their types (*"flavors"*) from electron neutrinos (to which the above detectors are sensitive) into other flavors that the detectors described above cannot see.

To understand these *neutrino flavor oscillations,* we must first understand

- weak interactions in the Standard Model of elementary particle physics and
- their properties in conjectured extensions of that model.

That will be the subject of following chapters.

10.5 Evolutionary Timescales

A question of basic importance is *how long a star will remain on the main sequence*.

- Evolution prior to the main sequence (protostar stage) is governed by *two primary timescales:*

  1. The *hydrodynamical timescale* (free-fall timescale)
  2. The *Kelvin–Helmholtz timescale* (thermal adjustment timescale).

  Evolution on the main sequence and beyond entails a *third timescale*, the *nuclear burning timescale.* For most stars:

  - The *hydrodynamical timescale is hours to days*.
  - The Kelvin–Helmholtz timescale is *hundreds of thousands to hundreds of millions of years*.
  - The nuclear burning timescale depends on the stellar fuel and mass, but is *much longer than the hydrodynamic and Kelvin–Helmholtz timescales*.

- Thus, stars spend much more time on the main sequence than in formation because

  > *Time spent on the main sequence is governed by the hydrogen burning timescale.*

  This timescale is much longer than the hydrodynamical and Kelvin–Helmholtz timescales.

For the Sun

- the hydrodynamical timescale is about *an hour*,
- the Kelvin–Helmholtz timescale is about *10 million years*, and
- the time to burn the core hydrogen fuel on the main sequence (nuclear burning timescale) is about *10 billion years.*

Once stars exhaust their core hydrogen and leave the main sequence, they can undergo *successive burnings of heavier fuels, which introduce new nuclear burning timescales*.

- In the periods between exhaustion of one fuel and ignition of another, thermal adjustment timescales will also be important.
- In certain cases (such as gravitational core collapse) hydrodynamical timescales will be relevant.
- Nuclear burning timescales after the main sequence are *longer than the corresponding Kelvin–Helmholtz and hydrodynamical timescales*, just as for the main sequence.
- But *post main-sequence burning timescales are much shorter than main-sequence hydrogen burning* because they occur at *much higher temperature and density*.
- Thus, a star generally spends more time on the main sequence than in its post main-sequence evolution.
- We conclude that

> *The nuclear burning timescale on the main sequence is longer than any other timescale in a star's life.*

- Thus at any one time *in a population of stars we expect to see the majority on the main sequence*
- (Unless the age is sufficiently large that most stars have had time to evolve off the main sequence).

10.6 Evolution of Stars on the Main Sequence

The main sequence is the *longest and most stable period* of a star's life, but

- *stars do evolve on the main sequence*, primarily in response to *core concentration changes* as they burn hydrogen to helium in hydrostatic equilibrium.
- This *lowers the pressure in the core* because it increases the mean molecular weight μ:

$$P = \rho \frac{kT}{\mu}$$

- This in turn
  - *increases the core density* and
  - releases *gravitational energy*,

  half of which is radiated away and half of which raises the core temperature (*virial theorem*).
- The energy outflow resulting from higher core temperatures *causes the outer layers to expand slightly and the star becomes more luminous*.
- Dependence of luminosity on the mean molecular weight μ is strong, varying as approximately $\mu^{7.5}$ (Exercise).

- The surface temperature during evolution on the main sequence may either increase or decrease.
  - For stars below about $1.25\,M_\odot$ the surface temperature tends to *increase*.
  - For more massive stars it tends to *decrease* as the star evolves on the main sequence.
- Therefore, the primary external effect of a star's evolution on the main sequence is to cause a small drift from the ZAMS position in the HR diagram:
  - Slightly upward and to the left for lighter stars.
  - Slightly upward and to the right for heavier stars.
- Internally the changes are more substantial, but their effects are often not very visible externally while the star continues to burn core hydrogen.
- Significant modification of elemental abundances is taking place as a result of the core fusion, but these changes are limited initially to the central regions.

The Standard Solar Model indicates that over the 4.6 billion year time that the Sun has spent on the main sequence

- The radius has increased by about 12%,
- The core temperature has increased by about 16%,
- The luminosity has increased by about 40%,
- The effective surface temperature has increased by about 3%, and the flux of ^{8}B neutrinos has increased by more than a factor of 40.
- Near the center the mass fraction of hydrogen has decreased and the mass fraction of helium has increased by about a factor of 2 from their initial values,
- Outside of about 20% of the solar radius hydrogen and helium retain their ZAMS abundances.
- The mass fraction of hydrogen fuel has decreased substantially in the solar core over its lifetime, but the rate of energy production by the PP chain is

  $$\frac{d\varepsilon}{dt} \simeq \rho^2 X^2 T^4,$$

  where ρ is the density, X the hydrogen mass fraction, and T the temperature.

  - Increasing ρ and T more than offset decreasing X as the Sun evolves on the main sequence.
  - This explains why *the Sun's luminosity is rising* even as its *hydrogen fuel is being depleted.*

Although the internal changes discussed in the preceding example lead to only small visible external modification of the star on the main sequence, *they set the stage for rapid evolution away from the main sequence* that will be the topic of subsequent chapters.

10.7 Timescale for Main Sequence Lifetimes

The rate of hydrogen fusion determines a timescale for life on the main sequence.

- Comparison of stellar evolution simulations with observations suggest that stars leave the main sequence when about *10% of their original hydrogen has been burned*.
- Let us define a timescale $\tau_{\rm nuc}$ for main sequence lifetimes by forming the ratio of
  - the *energy released from burning 10% of the hydrogen* and
  - the *luminosity*.
- The energy available from the burning of one gram of hydrogen to helium is $\sim 6 \times 10^{18}$ ergs. Therefore,

$$\tau_{\rm nuc} = \frac{E_{\rm H}/10}{L} = 6 \times 10^{17} \frac{XM}{L} \text{ s},$$

  - $E_{\rm H}$ is the energy available from fusing all the hydrogen in the star,
  - L is the present luminosity in erg s^{-1},
  - X is the original hydrogen mass fraction,
  - M is the mass of the star in grams.

Example: Inserting *values characteristic of the Sun*, we find for the Sun's main sequence timescale

$$\tau_{\mathrm{nuc}}^{\odot} = \frac{E_{\mathrm{H}}/10}{L} \sim 2.2 \times 10^{17}\,\mathrm{s} \sim 10^{10}\,\mathrm{yr}.$$

Expressing this timescale in solar units, we may write for any star

$$\tau_{\mathrm{nuc}} = 10^{10} \left(\frac{M}{M_{\odot}}\right) \left(\frac{L_{\odot}}{L}\right) \mathrm{yr},$$

and utilizing the mass–luminosity relation (Ch. 2)

$$\frac{L}{L_{\odot}} \simeq \left(\frac{M}{M_{\odot}}\right)^{3.5},$$

the main sequence timescale may be expressed for $M \geq M_{\odot}$ as

$$\tau_{\mathrm{nuc}} \simeq 10^{10} \left(\frac{M}{M_{\odot}}\right)^{-2.5} \mathrm{yr}.$$

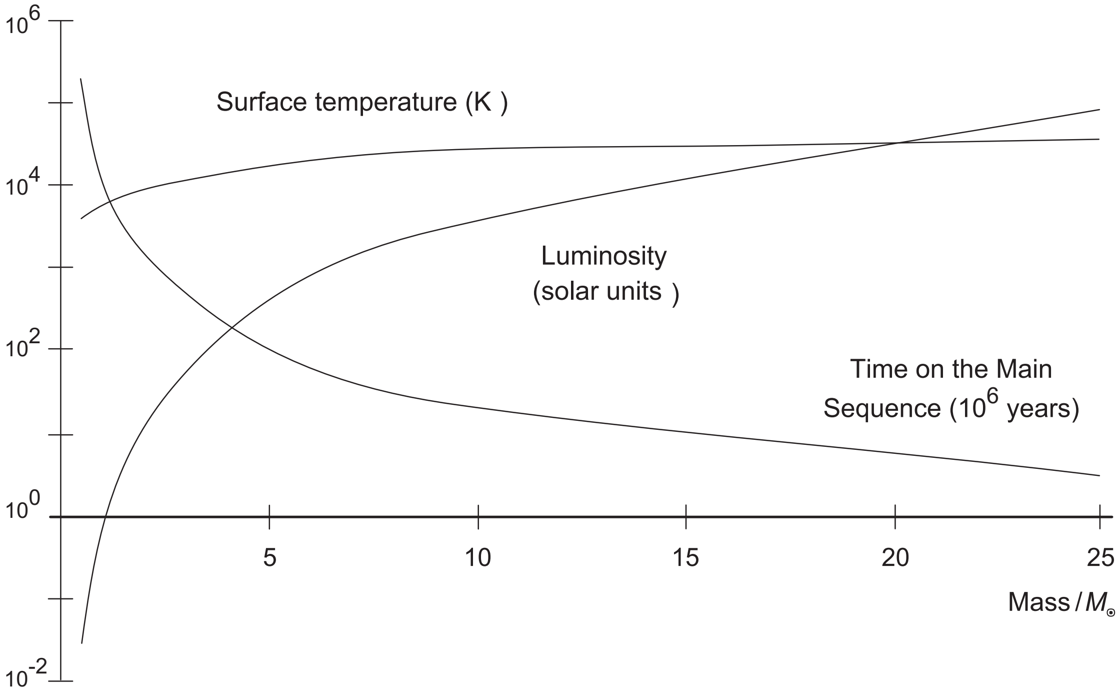


Figure 10.7: Main sequence lifetimes, temperatures, and luminosities.

Some main sequence lifetimes are illustrated in Fig. 10.7

- The Sun has a main sequence lifetime of about *10 billion years*, but
- a 20 $M_\odot$ star stays on the main sequence for about *5.5 million years* and
- a 100 $M_\odot$ star lives on the main sequence for only about *100,000 years*.
- Conversely, for main sequence stars with $M << M_\odot$, we may estimate that *the main sequence lifetime greatly exceeds the present age of the Universe*.

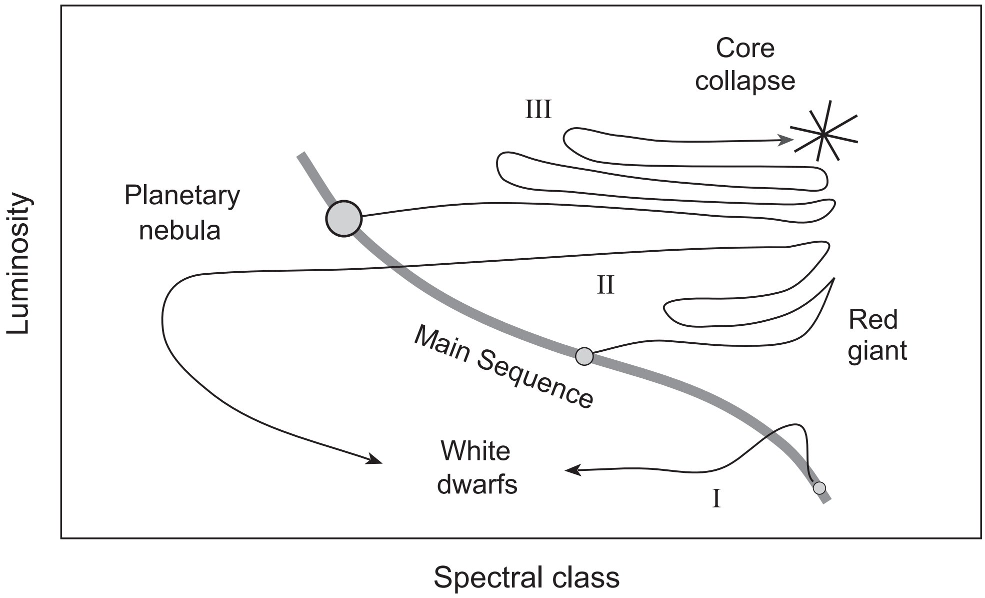


Figure 10.8: Categories of stellar evolution after the main sequence.

Depending on the mass of the star, evolution off the main sequence can lead to the *three qualitatively different scenarios* that are illustrated in Fig. 10.8.

- $M < 0.5M_\odot$: Core temperatures never rise high enough to ignite the He produced by PP-chain proton burning and *the star evolves to a He white dwarf*.
- $0.5 < M < 8M_\odot$: A red giant that eventually sheds much of its outer envelope as a planetary nebula and *becomes a C–O or Ne–Mg white dwarf*.
- $M > 8M_\odot$: A sequence of burning episodes involving successively heavier fuels until the core of the star collapses, producing in most cases a *supernova with a remnant neutron star or a black hole*.

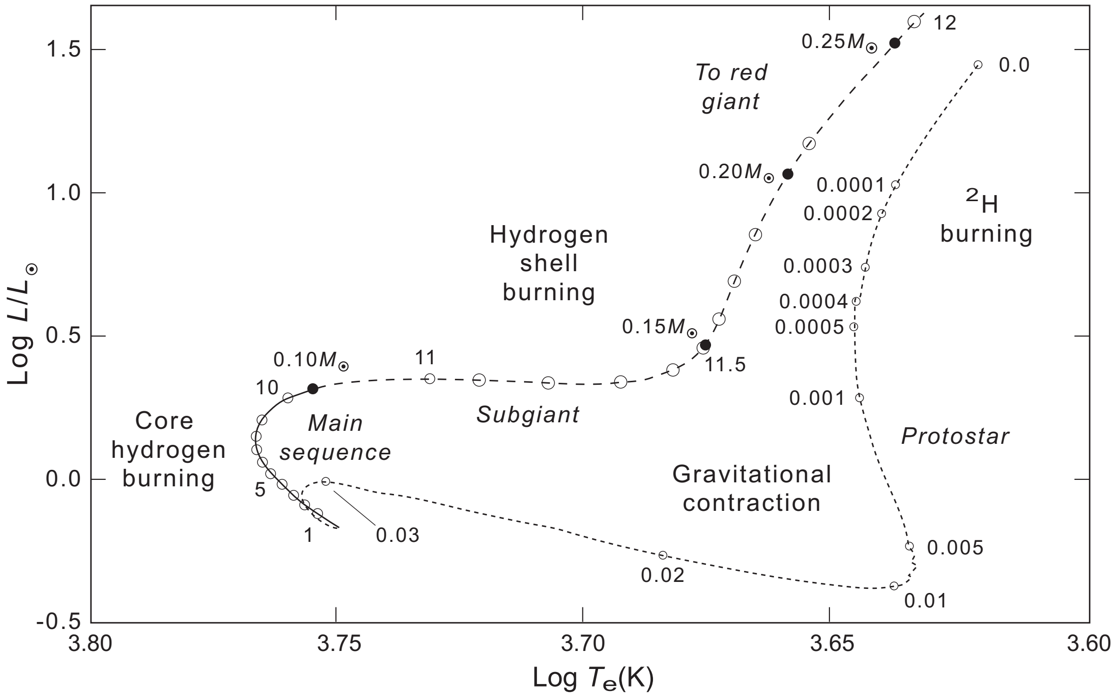


Figure 10.9: Simulated evolution of a 1 $M_\odot$ star from the final stages of protostar collapse through main-sequence core hydrogen burning and on to hydrogen shell burning. See text for explanation of symbols and number labels. Initial composition was $Y = 0.275$ and $Z = 0.015$. .

10.7.1 Examples of Post Main-Sequence Evolution

Figure 10.9 summarizes (1) the final stages of collapse to the main sequence, (2) evolution on the main sequence, and (3) evolution off the main sequence with the development of a shell hydrogen source for a $1M_\odot$ star.

- Numbers beside open circles indicate *times in* 10^9 *yr*.
- Numbers beside solid circles indicate *the mass coordinate for the hydrogen shell source*.

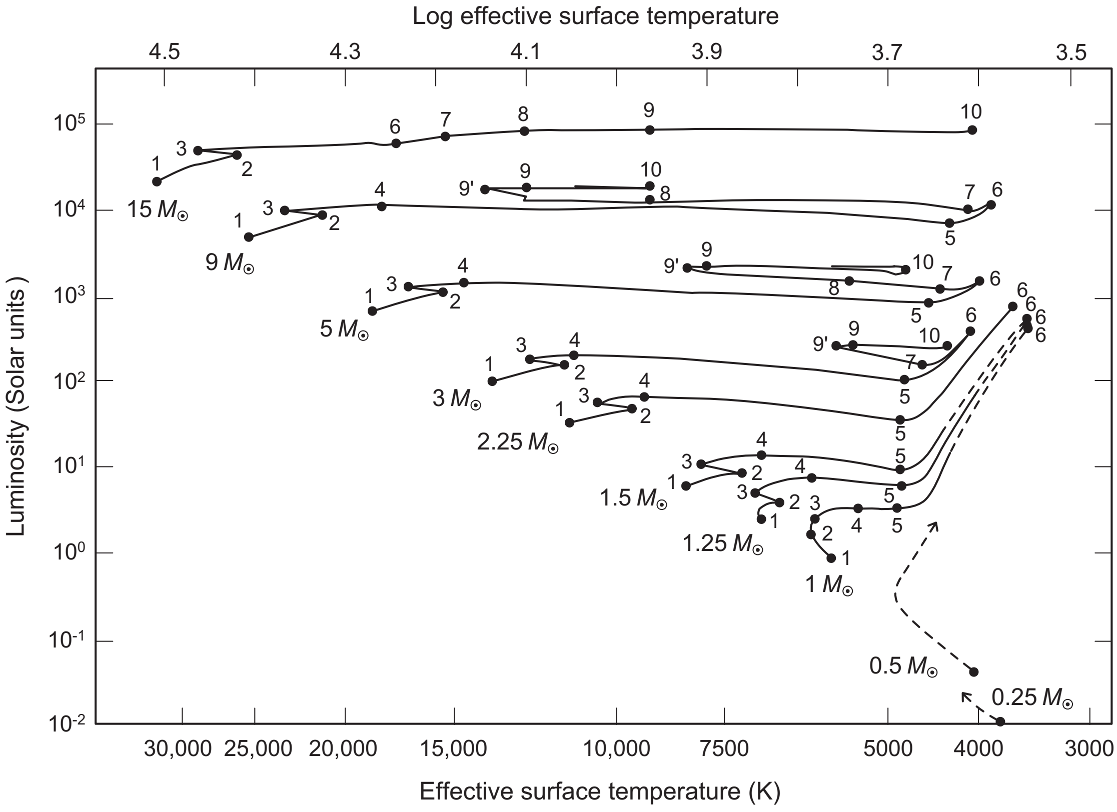


Figure 10.10: Evolution off the main sequence for stars of different initial main-sequence mass. Dashed lines are theoretical estimates.

- Evolution after the main sequence for other stars is qualitatively similar to the preceding examples, but the *details depend very much on the mass of the star*.
- An overview of initial evolution after leaving the main sequence for a range of initial main-sequence masses is given in the above figure.

Chapter 11

Flavor Oscillations of Solar Neutrinos

Recall from the previous chapter the solar neutrino problem:

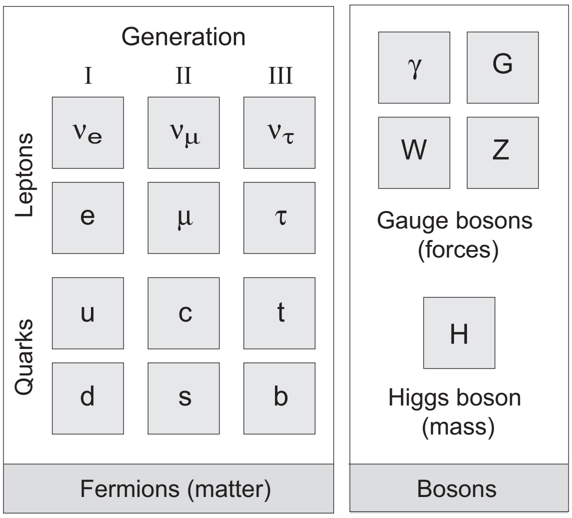
Solar neutrino fluxes from various experiments compared with a Standard Solar Model (SSM). All fluxes are in solar neutrino units (SNU), except the result from Super Kamiokande.

Experiment	Observed flux	SSM	Observed/SSM
Homestake	$2.54 \pm 0.14 \pm 0.14$ SNU	$9.3\,^{+1.2}_{-1.4}$	0.273 ± 0.021
SAGE	$72\,^{+12}_{-10}\,^{+5}_{-7}$ SNU	$137\,^{+8}_{-7}$	0.526 ± 0.089
GALLEX	$69.7 \pm 6.7\,^{+3.9}_{-4.5}$ SNU	$137\,^{+8}_{-7}$	0.509 ± 0.089
Super-Kamiokande	$2.51\,^{+0.14}_{-0.13}$ $(10^6\,\mathrm{cm}^{-2}\mathrm{s}^{-1})$	$6.62\,^{+0.93}_{-1.12}$	0.379 ± 0.034

In the preceding chapter we discussed the internal structure of the Sun and suggested that

- neutrinos emitted by thermonuclear processes in the central region of the Sun carry *direct information about the state of the core*.
- In this chapter we elaborate on the physics of solar neutrinos.
- Reconciliation of solar neutrino observations with our understanding of elementary particle physics has had *fundamental implications both for astrophysics and for elementary particle physics*.

Our discussion will involve directly only the Sun because it is the only normal star near enough to allow its neutrinos to be detected with present technology. However, *presumably the processes described in this chapter operate also in a similar way for other stars*.



Quark and neutrino mass scales			
Gen	I	II	III
Leptons	e^-, ν_e	μ^-, ν_μ	τ^-, ν_τ
Quarks	d, u	s, c	b, t
Quark mass (GeV)	~0.006	~1	~100
ν mass (GeV)	$<2.2\text{x}10^{-9}$	$<1.9\text{x}10^{-4}$	$<1.8\text{x}10^{-2}$

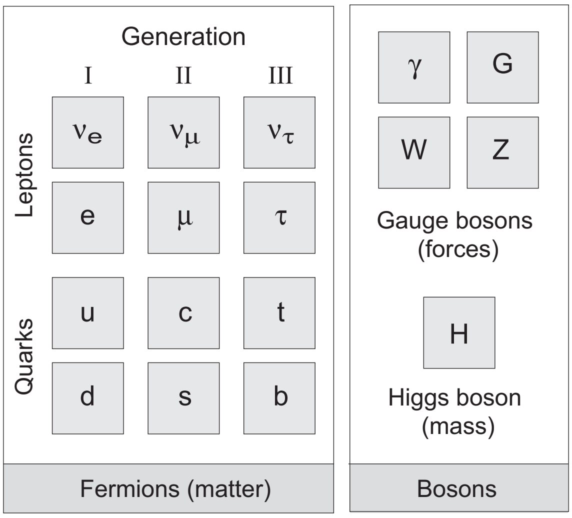
Figure 11.1: Particles of the Standard Model and characteristic mass scales in the quark and lepton sectors for each generation. Photons are labeled by γ and gluons by G.

11.1 Weak Interactions and Neutrino Physics

The *Standard Model* of elementary particle physics assumes that the electromagnetic and weak interactions are *unified in a local gauge or Yang–Mills theory* in which

- The leptons (electrons, neutrinos, . . .) and quarks are grouped into *generations or families* and
- They interact through the *exchange of gauge bosons: the photon, W^+, W^-, and Z^0*.

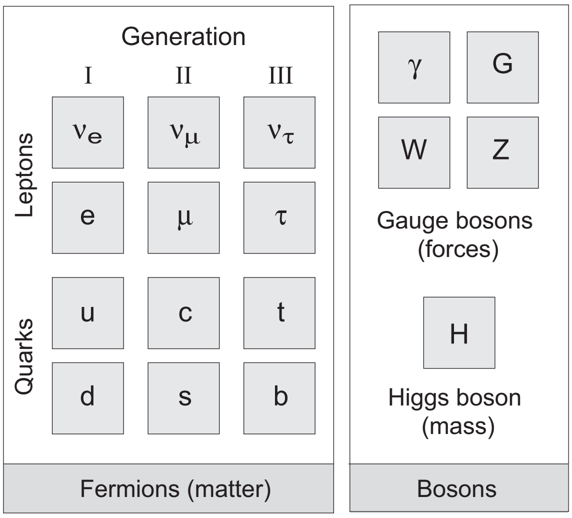
The particles of the Standard Model are listed in Fig. 11.1, along with their spins, charges, and masses (or experimental mass limits).



Quark and neutrino mass scales

Gen	I	II	III
Leptons	e^-, ν_e	μ^-, ν_μ	τ^-, ν_τ
Quarks	d, u	s, c	b, t
Quark mass (GeV)	~0.006	~1	~100
ν mass (GeV)	$<2.2\text{x}10^{-9}$	$<1.9\text{x}10^{-4}$	$<1.8\text{x}10^{-2}$

- In the Standard Model, the matter fields are divided into three “generations” or “families”, as illustrated in the figure above.
- An important ingredient of the Standard Model is that *the matter fields do not interact across family lines*.
- For the leptons this is implemented formally by assigning a lepton family number to each particle and requiring that interactions conserve this number.



Quark and neutrino mass scales

Gen	I	II	III
Leptons	e^-, ν_e	μ^-, ν_μ	τ^-, ν_τ
Quarks	d, u	s, c	b, t
Quark mass (GeV)	~0.006	~1	~100
ν mass (GeV)	$<2.2\text{x}10^{-9}$	$<1.9\text{x}10^{-4}$	$<1.8\text{x}10^{-2}$

Example: In Generation I

- Assign an electron family number of

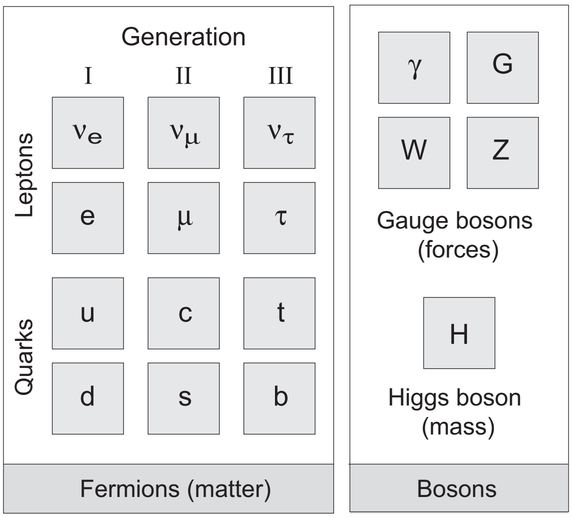
  (a) $+1$ to the electron and electron neutrino,

  (b) -1 to the antielectron and the electron antineutrino,

  (c) zero for all other particles.

- Then the reaction $\nu_e + n \to p + e^-$ *conserves electron family number* (because $1+0 = 0+1$) and is observed to occur.

- But the reaction $\nu_e + p \to n + e^+$ *violates electron family number* (because $1+0 \neq 0+(-1)$) and has never been observed.



Quark and neutrino mass scales

Gen	I	II	III
Leptons	e^-, ν_e	μ^-, ν_μ	τ^-, ν_τ
Quarks	d, u	s, c	b, t
Quark mass (GeV)	~0.006	~1	~100
ν mass (GeV)	$<2.2\text{x}10^{-9}$	$<1.9\text{x}10^{-4}$	$<1.8\text{x}10^{-2}$

- Also illustrated in the above table are *characteristic mass scales for quarks and neutrinos* within each generation.
- Neutrino masses are quoted as upper limits, since *no neutrino mass has been measured directly* thus far.
- The limits imply that *neutrino masses are either zero or tiny* on a mass scale set by the quarks of a generation.
- The explanation of this is a major unresolved issue in the theory of elementary particles.

In the *Standard Model* of elementary particle physics, it is *assumed* that the mass of all neutrinos (and antineutrinos) is identically zero, but some conjectured extensions allow a finite neutrino mass.

- The detailed explanation of this assumption requires quantum field theory beyond the scope of the present discussion.
- The central point is that there are potentially two kinds of neutrino mass terms that could appear in the theory:

  1. *Dirac masses* and
  2. *Majorana masses*.

- The first is appropriate if the *neutrino and antineutrino are separate particles*.
- The second is appropriate if *the neutrino is its own antiparticle* (which is not ruled out by present data).
- *Both types of mass terms must vanish for Standard Model neutrinos,* but at least one could be non-zero in various extensions of the Standard Model.

11.1.1 Charged and Neutral Weak Currents

The Standard Model permits two basic classes of weak interactions.

- In *charged weak currents* electrical charge is transferred in the interaction
  - The *total charge is conserved*, of course.
  - Because charge is transferred, *the boson mediating the force must be charged*.
  - Thus, *charged weak currents involve the* W^+ *and* W^- weak gauge bosons.
- The uncharged weak gauge boson Z^0 can mediate *neutral weak currents* in which there is *no transfer of charge* in the weak interaction matrix elements.

A $\nu_{\rm e}$ can interact with an electron through the charged weak current or the neutral weak current. However,

- A ν_μ or ν_τ cannot interact with an electron through exchange of a charged gauge boson without violating lepton family number conservation.
- Therefore, *only* $\nu_{\rm e}$ *can undergo charged-current interactions with electrons*.

In contrast, *neutral current interactions can take place with any flavor of neutrino* without violating lepton family number conservation.

- Thus, electron neutrinos can interact with the electrons in normal matter through *both the charged and neutral weak currents*.
- All other flavors of neutrinos can interact with electrons *only through the neutral weak current*.

11.1.2 Mixing in the Quark Sector

The term *flavor* is used to distinguish the quarks and leptons of one generation from another.

- We shall often refer to ν_{e}, ν_μ, and ν_τ as *different neutrino flavors*, and to u, d, s, ... as *different quark flavors*.
- In the *Standard Model*, observations require that quark mass eigenstates and weak eigenstates are *not equivalent:*.

Quark states entering the weak interactions are *linear combinations of mass eigenstates*.

- *Example:* The d and s quarks enter the weak interactions in the *"rotated" linear combinations d_{c} and s_{c}* defined by

$$\underbrace{\begin{pmatrix} d_{\mathrm{c}} \\ s_{\mathrm{c}} \end{pmatrix}}_{\text{weak eigenstates}} = \underbrace{\begin{pmatrix} \cos\theta_{\mathrm{c}} & \sin\theta_{\mathrm{c}} \\ -\sin\theta_{\mathrm{c}} & \cos\theta_{\mathrm{c}} \end{pmatrix}}_{\text{flavor mixing matrix}} \underbrace{\begin{pmatrix} d \\ s \end{pmatrix}}_{\text{mass eigenstates}}$$

  where in this matrix equation

  - $d \equiv |d\rangle$ and $s \equiv |s\rangle$ are *mass-eigenstate quark fields*,
  - θ_{c} is termed the *mixing angle* or the *Cabibbo angle*.
  - $|x\rangle$ is the *quantum wavefunction* of quark field x.

This matrix equation is a compact way to write *two equations:*

$$|d_{\mathrm{c}}\rangle = \cos\theta_{\mathrm{c}}\,|d\rangle + \sin\theta_{\mathrm{c}}\,|s\rangle\,,$$
$$|s_{\mathrm{c}}\rangle = -\sin\theta_{\mathrm{c}}\,|d\rangle + \cos\theta_{\mathrm{c}}\,|s\rangle\,.$$

- Comparison with data reveals that *the Cabbibo mixing angle is small:*

$$\sin\theta_c \simeq 0.230 \qquad \cos\theta_c \simeq 0.973.$$

- In the realistic case of *three generations of quarks*, weak eigenstates are described by a 3 × 3 mixing matrix called the *Cabibbo–Kobayashi–Maskawa or CKM matrix* that is parameterized by
  - three real mixing angles
  - one phase.

There is little fundamental understanding of this quark flavor mixing but the data require it.

11.1.3 Mixing in the Leptonic Sector

In the Standard Model the quarks entering the weak interactions are mixtures of different mass eigenstates.

- Hence we might expect that the corresponding leptons in these generations could also enter the weak interactions as mixed mass eigenstates.
- *If all flavors of neutrinos are identically massless* (as is assumed in the Standard Model), *flavor mixing has no observable consequences.*
- However, if at least one neutrino has a non-zero mass, neutrino flavor mixing could have observable consequences.
- Conversely, *observation of neutrino flavor mixing* is a direct indication that *at least one neutrino has a finite mass.*

Thus either the

- observation of *neutrino flavor oscillations*, or
- direct measurement of a *finite neutrino mass*

would imply the existence of *physics beyond the Standard Model*.

11.2 Beyond the Standard Model: Finite Neutrino Masses

As noted above, a *finite neutrino mass implies possible flavor mixing*.

- This is of *fundamental importance for elementary particle physics* because it would imply new physics beyond the Standard Model.
- But it could be of equal importance for astrophysics because it provides a *possible solution of the solar neutrino problem:*
  - As we shall show below, if neutrino flavor eigenstates are mixtures of mass eigenstates, *neutrinos propagating in time will oscillate in flavor*.
  - Then it is possible that when some of the electron neutrinos emitted by the Sun reach Earth they would have *oscillated into another flavor.*
  - Since the experiments described earlier are *sensitive only to electron neutrinos,* they would miss any neutrinos that had oscillated into other flavors.
  - This could (possibly) explain the observed neutrino deficit.

Standard Model neutrinos *must* be massless.

- However, there are many reasons to believe that the Standard Model—despite its remarkable success—is incomplete and represents a low-energy approximation to a more complete theory.
  - There are $\sim$20 adjustable parameters that have no convincing fundamental constraint.
  - The origin of mass through the Higgs mechanism is purely phenomenological.
  - The generational (family) structure is based entirely on phenomenology.
  - Violations of symmetries such as parity are put by hand, . . .
- Various extensions such as *Grand Unified Theories (GUTs)* have been proposed that go beyond the Standard Model.
- For these theories often the reasons that mass terms are forbidden in the Standard Model are not operative and *neutrino mass terms may occur naturally.*

We must entertain the possibility of physics beyond the Standard Model and thus of finite neutrino masses.

11.3 Neutrino Vacuum Oscillations

The preceding discussion suggests that neither

- direct experimental measurement,
- nor fundamental principle,
- nor our present understanding of the Standard Model extended to Grand Unified Theories

preclude a small mass for neutrinos.

Therefore, let us pursue the possibility that finite-mass neutrinos undergo flavor oscillations, and that these oscillations could account for the solar neutrino deficit.

11.3.1 Mixing for Two Neutrino Flavors

Consider neutrino oscillations in a two-flavor model, in the absence of matter. (Note: $\hbar = c = 1$ units)

- We shall term these *vacuum oscillations*, as opposed to oscillations that occur in matter.
- Generally, we shall use θ to denote neutrino vacuum oscillation angles and θ_m to denote oscillation angles in matter.
- The *flavor (weak) eigenstates* $|\nu_\mathrm{e}\rangle$ and $|\nu_\mu\rangle$ may be expressed in terms of the mass eigenstates $|\nu_1\rangle$ and $|\nu_2\rangle$ through the matrix transformation

$$\underbrace{\begin{pmatrix} \nu_\mathrm{e} \\ \nu_\mu \end{pmatrix}}_{\text{flavor eigenstates}} = \underbrace{\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}}_{\text{flavor mixing matrix}} \underbrace{\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}}_{\text{mass eigenstates}},$$

where θ is the *vacuum mixing angle.* Thus,

$$|\nu_\mathrm{e}\rangle = \cos\theta_\mathrm{v}\,|\nu_1\rangle + \sin\theta_\mathrm{v}\,|\nu_2\rangle \qquad |\nu_\mu\rangle = -\sin\theta_\mathrm{v}\,|\nu_1\rangle + \cos\theta_\mathrm{v}\,|\nu_2\rangle\,.$$

- By *inverting this expression*, mass eigenstates may in turn be expressed as a linear combination of flavor eigenstates:

$$\underbrace{\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}}_{\text{mass eigenstates}} = \underbrace{\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}}_{\text{flavor mixing matrix}} \underbrace{\begin{pmatrix} \nu_\mathrm{e} \\ \nu_\mu \end{pmatrix}}_{\text{flavor eigenstates}}.$$

- Assuming that *at least one of the neutrino masses is non-zero*, the different mass eigenstates will move with *slightly different speeds* as neutrinos propagate in time.
- Thus, the expansion coefficients in the above equation will vary with time and
- the *probability of detecting a particular flavor of neutrino will oscillate with time*, or equivalently, with the distance traveled.
- From basic quantum field theory the *mass eigenstates* for neutrinos evolve with time t according to

  $$|\nu_i(t)\rangle = e^{-iE_it}\,|\nu_i(0)\rangle\,,$$

  where the index i labels mass eigenstates of energy E_i.
- Thus the time evolution of the $\nu_{\text{e}} = \cos\theta_{\text{v}}\,|\nu_1\rangle + \sin\theta_{\text{v}}\,|\nu_2\rangle$ state will be given by

  $$\begin{aligned}|\nu(t)\rangle &= \cos\theta_{\text{v}}\,|\nu_1(t)\rangle + \sin\theta_{\text{v}}\,|\nu_2(t)\rangle \\ &= \cos\theta e^{-iE_1t}\,|\nu_1(0)\rangle + \sin\theta e^{-iE_2t}\,|\nu_2(0)\rangle\,,\end{aligned}$$

  and this may be expressed as the mixed-flavor state

  $$\begin{aligned}|\nu(t)\rangle = (&\cos^2\theta e^{-iE_1t} + \sin^2\theta e^{-iE_2t})\,|\nu_{\text{e}}\rangle \\ &+ \sin\theta\cos\theta(-e^{-iE_1t} + e^{-iE_2t})\,\big|\nu_\mu\big\rangle\,.\end{aligned}$$

- We see that the mixed flavor state

$$|\nu(t)\rangle = (\cos^2\theta e^{-iE_1t} + \sin^2\theta e^{-iE_2t})\,|\nu_{\rm e}\rangle \\ + \sin\theta\cos\theta(-e^{-iE_1t} + e^{-iE_2t})\,\big|\nu_\mu\big\rangle .$$

starts out as *pure* $\nu_{\rm e}$ *at* $t = 0$,

$$|\nu(0)\rangle = \underbrace{(\cos^2\theta + \sin^2\theta)}_{=1}|\nu_{\rm e}\rangle + \underbrace{\sin\theta\cos\theta(-1+1)}_{=0}\big|\nu_\mu\big\rangle = |\nu_{\rm e}\rangle ,$$

- but will be *mixed* $\nu_{\rm e}$ *and* ν_μ *after a finite time*.

- Taking the overlap $\langle \nu_i|\,\nu(t)\rangle$ of the time-evolving mixed-flavor state

$$|\nu(t)\rangle = (\cos^2\theta e^{-iE_1t} + \sin^2\theta e^{-iE_2t})\,|\nu_{\rm e}\rangle \\ + \sin\theta\cos\theta(-e^{-iE_1t} + e^{-iE_2t})\,\big|\nu_\mu\big\rangle .$$

with the flavor eigenstates $|\nu_i\rangle$, we find that the *probabilities for an initial electron neutrino state to remain an electron neutrino, or be converted to a muon neutrino after a time* t, are given by

$$\begin{aligned} P(\nu_{\rm e} \to \nu_{\rm e}, t) &= |\langle \nu_{\rm e}|\,\nu(t)\rangle|^2 \\ &= 1 - \tfrac{1}{2}\sin^2(2\theta)[1 - \cos(E_2 - E_1)t] \qquad \text{(remain } \nu_{\rm e}) \\ P(\nu_{\rm e} \to \nu_\mu, t) &= |\big\langle \nu_\mu\big|\,\nu(t)\rangle|^2 \\ &= \tfrac{1}{2}\sin^2(2\theta)[1 - \cos(E_2 - E_1)t] \qquad \text{(become } \nu_\mu). \end{aligned}$$

with the *sum of probabilities equal to one*.

11.3.2 The Vacuum Oscillation Length

- Assuming the energies E_1 and E_2 to be *approximately equal and much larger than* mc^2 for the neutrino masses,

$$E_i = (p^2 + \underbrace{m_i^2}_{\text{small}})^{1/2} \simeq \underbrace{p + \frac{m_i^2}{2p}}_{\text{binomial exp.}} \longrightarrow \quad E_2 - E_1 \simeq \frac{\Delta m^2}{2E}$$

  where $E_1 \sim E_2 \equiv E$ and $\Delta m^2 \equiv m_2^2 - m_1^2$.

- Probabilities for *flavor survival and conversion* as a function of *distance traveled* $r \sim ct$ may then be expressed as

$$P(\nu_{\rm e} \to \nu_{\rm e}, r) = 1 - \sin^2(2\theta)\sin^2\left(\frac{\pi r}{L}\right),$$
$$P(\nu_{\rm e} \to \nu_\mu, r) = \sin^2(2\theta)\sin^2\left(\frac{\pi r}{L}\right),$$

  where the *oscillation length* L is defined by

$$L \equiv \frac{4\pi E}{\Delta m^2}.$$

  Physically L is the *distance required for one one complete flavor oscillation* (for example, $\nu_{\rm e} \to \nu_\mu \to \nu_{\rm e}$).

- Restoring the factors of $\hbar$ and c in the preceding equation

$$L = \frac{4\pi E\hbar}{\Delta m^2 c^3} = 2.48\left(\frac{E}{\rm MeV}\right)\left(\frac{\rm eV^2}{\Delta m^2}\right),$$

  where L is in meters, the neutrino energy E is in MeV, and the mass squared difference Δm^2 is in eV^2.

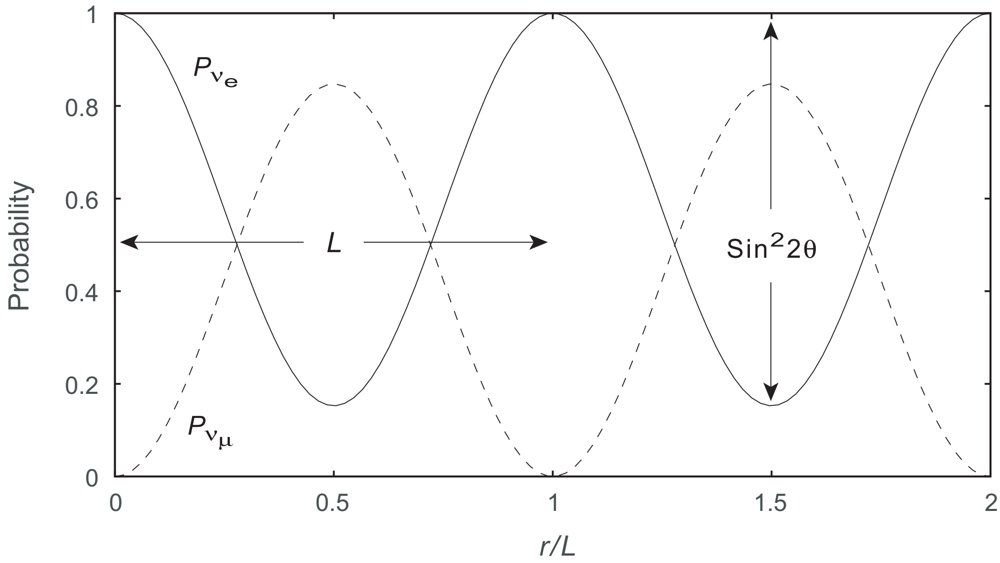


Figure 11.2: Neutrino vacuum oscillations in a 2-flavor model as a function of distance traveled r in units of the vacuum oscillation length L. The probability as a function of r to be an electron neutrino is denoted by P_{ν_e} and that to be a muon neutrino by P_{ν_μ}. The period of the oscillation is L and its amplitude is $\sin^2 2\theta$, where θ is the vacuum mixing angle. In this calculation $\theta = 33.5°$, the neutrino energy is $E = 5\,\mathrm{MeV}$, the difference in squared masses for the two flavors is $\Delta m^2 c^4 = 7.5 \times 10^{-5}\,\mathrm{eV}^2$, and the corresponding oscillation length L is 165.3 km.

Neutrino oscillations for a 2-flavor model using these formulas are illustrated in Fig. 11.2.

11.3.3 Time-Averaged or Classical Probabilities

The oscillation wavelength may be smaller than the uncertainties in position for emission and detection of neutrinos.

- There are thousands of kilometers variation in the distance between production and detection of solar neutrinos due to
  - *varying production location* in the Sun,
  - *varying detection location* since the Earth is rotating,
  - and *varying Earth–Sun separation*.
- If the oscillation length is less than the averaging introduced by the preceding considerations, the detectors will see a *distance (or time) average*.
- Denoting the averaged detection probability by a bar gives

$$\bar{P}(\nu_{\rm e} \to \nu_{\rm e}) = 1 - \tfrac{1}{2}\sin^2 2\theta \qquad \bar{P}(\nu_{\rm e} \to \nu_{\mu}) = \tfrac{1}{2}\sin^2 2\theta .$$

- The average survival probability has a *lower limit of* $\frac{1}{2}$ for two flavors.
- For n flavors the lower limit is n^{-1}, but that limit can be realized only for a precisely-tuned flavor mixture.
- The above results are equivalent to the result obtained if the interference terms resulting from squaring amplitudes in quantum mechanics are discarded.
- Thus the *average probability* may also be viewed as the *classical probability*.

For later use we note that the classical flavor-conversion probability can be written as a matrix equation:

- Letting the row vector $(1\ \ 0)$ and its corresponding column vector denote pure $\nu_{\rm e}$ flavor states,

$$\bar{P}(\nu_{\rm e} \to \nu_{\rm e}) = \sum_{i=1}^{2} P(\nu_{\rm e} \to \nu_i)P(\nu_i \to \nu_{\rm e}) =$$

$$(1\ \ 0)\begin{pmatrix} \cos^2\theta & \sin^2\theta \\ \sin^2\theta & \cos^2\theta \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \cos^2\theta & \sin^2\theta \\ \sin^2\theta & \cos^2\theta \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

- which you can verify by matrix multiplication and trigonometric identities is equivalent to the earlier result

$$\bar{P}(\nu_{\rm e} \to \nu_{\rm e}) = 1 - \tfrac{1}{2}\sin^2 2\theta$$

for the classical probability to remain a $\nu_{\rm e}$.

11.4 Neutrino Oscillations with Three Flavors

We will demonstrate in the next chapter that solar neutrinos may be understood well in the simple 2-flavor formalism developed above. Thus at fixed energy a single vacuum mixing angle and one mass-squared difference characterizes the theory. However, in the general case there are three known flavors of neutrinos (and their corresponding antineutrinos), so the correct treatment of neutrino oscillations requires additional parameters associated with a 3×3 mixing matrix. We refer to the book for the resulting mixing matrix.

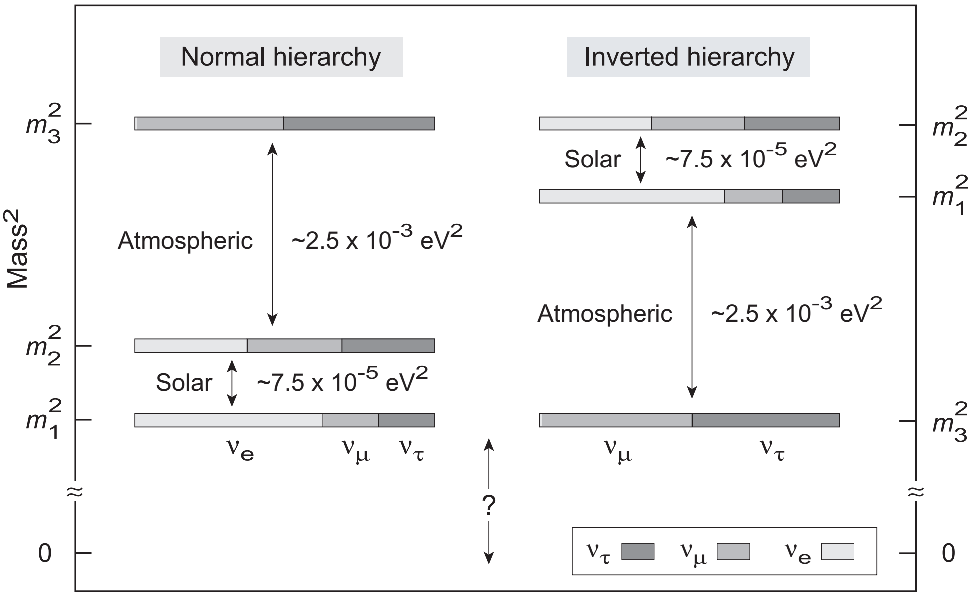


Figure 11.3: Neutrino mass hierarchy in a three-flavor model. Mass-square differences inferred from atmospheric neutrino and solar neutrino data are indicated. Since only values of Δm^2 and not absolute masses are known, two orderings of the known mass square differences are consistent with data: the *normal hierarchy* and the *inverted hierarchy*. Shading indicates the relative contribution of the three neutrino flavors to each mass eigenstate in the two possible orderings.

11.4.1 The Neutrino Mass Hierarchy

Direct mass measurements place only upper limits on neutrino masses for the three flavors. Oscillation measurements indicate

- that at least *some neutrinos flavors have finite mass*, and
- constrain the mixing angle and the mass squared difference between flavors, but *cannot give the actual masses*.

This leads to the *hierarchy ambiguity* displayed in Fig. 11.3.

In principle the correct hierarchy can be inferred from matter oscillations but evidence for these has been seen thus far

- only for solar neutrinos and
- not for atmospheric neutrinos (described in the next chapter).

Thus only *mass differences* are known presently for neutrinos.

Chapter 12

Solar Neutrinos and the MSW Effect

The vacuum neutrino oscillations described in the previous section could in principle account for the depressed flux of solar neutrinos detected on Earth.

- But this solution requires a *large mixing angle* to suppress the electron neutrino flux sufficiently.
- *Quark-sector mixing angles are relatively small,* so theoretical prejudice (but no evidence) initially favored a small mixing angle in the neutrino sector.
- However, there is another issue: *the neutrinos also have to transit out of the Sun.*
- *Electron neutrinos couple more strongly to normal matter than do other neutrinos* (because electron neutrinos and the particles of normal matter all reside in the first generation of the Standard Model).
- Thus we must also ask how interaction with solar material will influence neutrino oscillations.

12.1 The Mass Matrix

Let's first introduce an alternative formulation for 2 flavors.

- A *neutrino propagating in vacuum* can be written as a time-dependent linear combination of flavor eigenstates

  $$|\nu(t)\rangle = \nu_{\mathrm{e}}(t)\,|\nu_{\mathrm{e}}\rangle + \nu_{\mu}(t)\,|\nu_{\mu}\rangle\,,$$

  where $\nu_{\mathrm{e}}(t)$ and $\nu_{\mu}(t)$ obey the matrix equation

  $$i\frac{d}{dt}\begin{pmatrix} \nu_{\mathrm{e}}(t) \\ \nu_{\mu}(t) \end{pmatrix} = M_0 \begin{pmatrix} \nu_{\mathrm{e}}(t) \\ \nu_{\mu}(t) \end{pmatrix}.$$

- The *mass matrix in vacuum*, M_0, is given by

  $$M_0 = \begin{pmatrix} E_1\cos^2\theta + E_2\sin^2\theta & (E_2 - E_1)\sin\theta\cos\theta \\ (E_1 - E_2)\sin\theta\cos\theta & E_1\sin^2\theta + E_2\cos^2\theta \end{pmatrix},$$

  where θ is the mixing angle and

  $$E_i = (p^2 + m_i^2)^{1/2} \simeq p + \frac{m_i^2}{2p},$$

  assuming that $E_i >> m_i c^2$.

- After *subtracting a multiple of the unit matrix* (no influence on flavor probabilities), the vacuum mass matrix is

  $$M_0 = \frac{\pi}{L}\begin{pmatrix} \cos 2\theta & -\sin 2\theta \\ -\sin 2\theta & -\cos 2\theta \end{pmatrix},$$

  where L is the *vacuum oscillation length*.

12.2 Propagation of Neutrinos in Matter

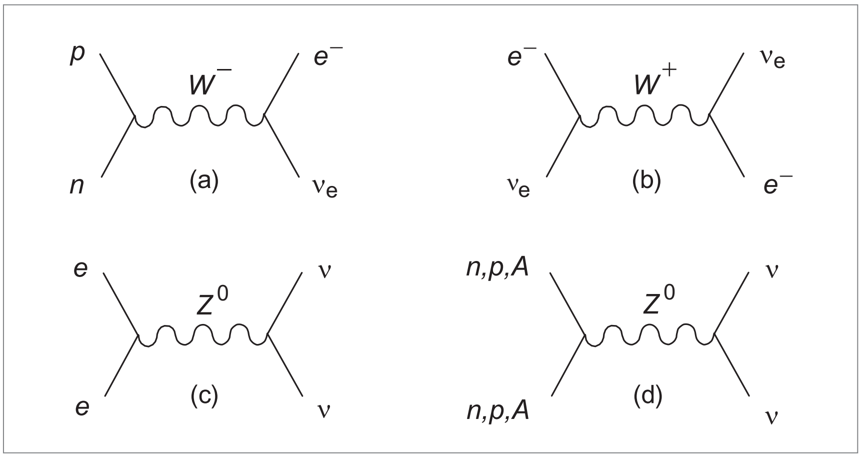
So far, this is just a reformulation of our previous equations for vacuum oscillation.

- Now, following the insight of *Mikheyev, Smirnov, and Wolfenstein (MSW),* we consider the additional influence that *interaction with matter* may have on the neutrino oscillation.
- When electron neutrinos scatter elastically from electrons in the Sun, they may do so through either
  - the charged weak current or
  - the neutral weak current.

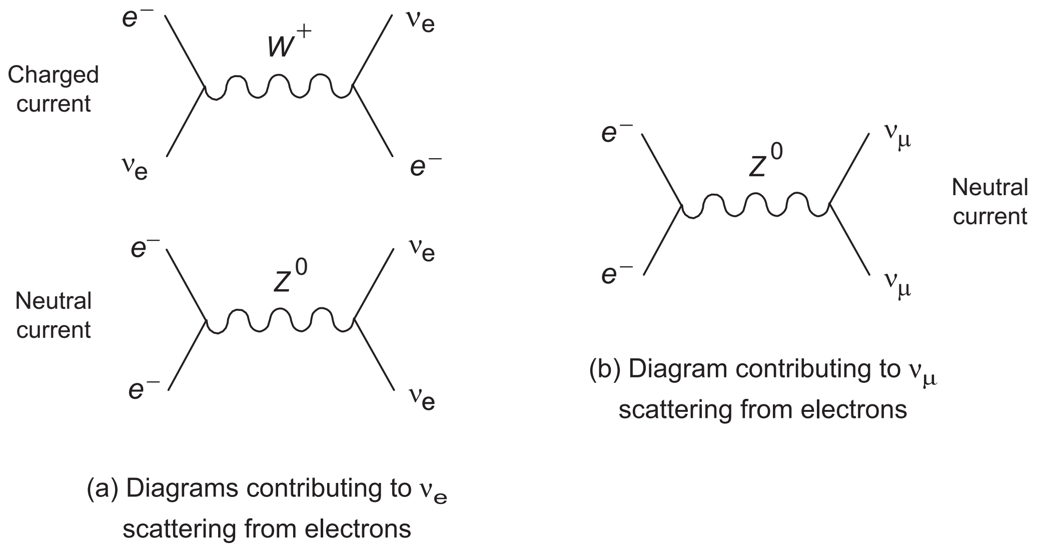
Feynman Diagrams

In quantum field theory we use a pictorial representation of interaction matrix elements called *Feynman diagrams.*

- They are highly intuitive: given a Feynman diagram one can write the corresponding matrix element and given the matrix element one can sketch the Feynman diagram.
- Here are some weak-interaction Feynman diagrams:



- The solid lines represent (fermion) matter fields and the wiggly lines represent exchanged virtual gauge bosons.
- Each diagram can represent several related processes.
- For example, diagram (a) read from the bottom:
  1. A neutron (n) exchanges a virtual W^- intermediate vector boson with an electron neutrino (ν_e).
  2. This converts $n \longrightarrow p$ and $\nu_\mathrm{e} \longrightarrow e^-$.
- Absence of flavor indices on neutrinos in diagrams (c) and (d) indicates that the neutral current is flavor blind. The symbol A in diagram (d) stands for a composite nucleus.



(a) Diagrams contributing to ν_e scattering from electrons

(b) Diagram contributing to ν_μ scattering from electrons

Figure 12.1: Feynman diagrams responsible for neutrino–electron scattering in the MSW effect.

Figure 12.1 illustrates Feynman diagrams relevant to neutrino scattering in the Sun.

- Both the neutral and charged current contribute to electron neutrino interactions (left two diagrams).
- Only the neutral current contributes to the muon neutrino interactions.
- Neutral current contributes to both ν_e and ν_μ scattering, so neglect this common contribution for this discussion.
- Vacuum neutrino oscillations will be modified in matter because of the charged-current (W^+) diagram in Fig. 12.1 contributing to ν_e scattering but not to ν_μ.

The situation is similar to two coupled oscillators where the frequency of one oscillator is modified more by coupling to its surroundings than the other. Such a modification will influence the nature of the coupling between the two oscillators.

- The *charged-current Feynman diagram* contributes an additional term to the vacuum-scattering mass matrix that may be expressed as

$$V = \sqrt{2} G_\mathrm{F} n_\mathrm{e},$$

  where G_F is the weak coupling constant, n_e is the local electron number density, and

- The potential V is *seen only by electron neutrinos*.

- With this additional contribution, the *mass matrix in the presence of matter* becomes

$$M = \frac{\pi}{L} \begin{pmatrix} \cos 2\theta + L/\ell_\mathrm{m} & -\sin 2\theta \\ -\sin 2\theta & -\cos 2\theta - L/\ell_\mathrm{m} \end{pmatrix}.$$

- The additional matter contribution to the oscillation length ℓ_m is given by

$$\ell_\mathrm{m} = \frac{\sqrt{2}\pi}{G_\mathrm{F} n_\mathrm{e}}.$$

12.2.1 The Effective Neutrino Mass in Medium

For the electron neutrino subject to the additional potential V we have

$$E - V = (p^2 + m^2)^{1/2}$$

and hence

$$p^2 + m^2 = (E-V)^2 = E^2 \left(1 - \frac{V}{E}\right)^2 \simeq E^2 - 2EV,$$

where the last step is justified by the assumption that $V \ll E$. Thus the energy of the electron neutrino propagating in the medium is

$$E^2 \sim p^2 + \tilde{m}^2 \qquad \tilde{m} \equiv (m^2 + 2EV)^{1/2}$$

where

- $\tilde{m}$ may be interpreted as an *effective mass* that has been modified from its value in vacuum by interaction with the medium.
- Since V is positive, an electron neutrino behaves effectively as if it is slightly heavier when propagating through matter than in vacuum,
- with the amount of mass increase governed by the electron number density of the matter.
- Fig. 12.2 on the following page illustrates.

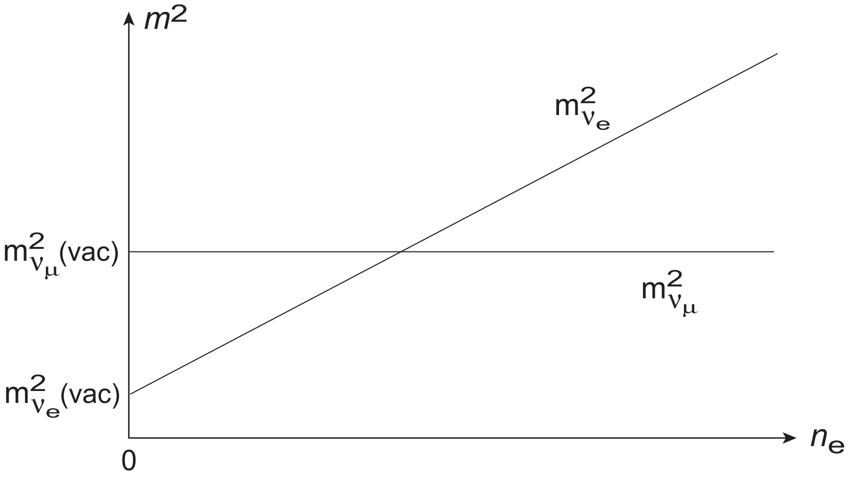


Figure 12.2: Effective mass-squared of electron neutrinos and muon neutrinos as a function of electron number density n_e, neglecting flavor mixing. Because the ν_μ does not couple to the charged weak current its m^2 does not depend on n_e but the effective m^2 of ν_e increases linearly with the electron density. The order of states in the m^2 spectrum in vacuum (left side) can become *inverted in matter* (right side).

Fig. 12.2 shows that the charged-current changes the effective mass of an electron neutrino in medium.

- An electron neutrino less massive than a muon neutrino in vacuum will become effectively *more massive* in matter if the electron density is high enough: *the m^2 spectrum can become inverted.*
- Gaining an effective mass through interaction with a medium is common for particles in many contexts.

Example: A superconductor expels a magnetic field because a photon gains an effective mass inside it (*Meissner effect*).

12.2.2 Propagation of Left-Handed Neutrinos

Only the left-handed component of a neutrino couples to the weak interactions.

- Thus for $E \gg m$ only the propagation of left-handed neutrinos is relevant.

The Schrödinger equation of ordinary quantum mechanics is not relativistically invariant.

- For relativistic fermions the wave equation must be generalized to the *Dirac equation*, while
- for spinless particles the corresponding relativistic wave equation is the *Klein–Gordon equation.*

Neutrinos are ultrarelativistic fermions but

- the propagation of just the left-handed component of the free neutrino may be described by the simple *free-particle Klein–Gordon equation,*

$$(\Box + m^2)\,|\nu\rangle = 0 \qquad \Box \equiv -\frac{\partial^2}{\partial t^2} + \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

- where $\Box$ is termed the *d'Alembertian operator* and
- for n neutrino flavors $|\nu\rangle$ is an n-component column vector in the mass-eigenstate basis and
- m^2 is an $n \times n$ matrix.

Because of oscillations, the solutions of interest correspond to the propagation of a linear combination of mass eigenstates.

- For ultrarelativistic neutrinos we make only small errors by assuming neutrinos of

  - tiny mass and
  - slightly different energies

  to propagate with the same 3-momentum p.

- In that approximation a solution of the Klein–Gordon equation for definite momentum is given by

$$|\nu_i\rangle = e^{-iE_it} \cdot e^{-ip\cdot x} \qquad E_i = \sqrt{p^2 + m_i^2}.$$

For ultrarelativistic particles this may be approximated as

$$|\nu_i(t)\rangle \simeq e^{-i(m_i^2/2E)t}.$$

Differentiating with respect to time gives an equation of motion for a single mass eigenstate

$$i\frac{d}{dt}\,|\nu_i(t)\rangle = \frac{m_i^2}{2E}\,|\nu_i(t)\rangle\,,$$

which may be generalized for a 2-flavor model to the matrix equation

$$i\frac{d}{dt}\begin{pmatrix}\nu_1\\ \nu_2\end{pmatrix} = M\begin{pmatrix}\nu_1\\ \nu_2\end{pmatrix} = \begin{pmatrix}m_1^2/2E & 0\\ 0 & m_2^2/2E\end{pmatrix}\begin{pmatrix}\nu_1\\ \nu_2\end{pmatrix},$$

where M is termed the *mass matrix.*

12.2.3 Evolution in the Flavor Basis

Neutrinos propagate in mass eigenstates but they are produced and detected in flavor eigenstates,

- so it is useful to express the preceding in the flavor basis.
- The required transformations are given by

$$\begin{pmatrix} \nu_{\rm e} \\ \nu_\mu \end{pmatrix} = U \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$$

$$U = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \qquad U^\dagger = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

permitting the evolution equation to be written in the form

$$i\frac{d}{dt}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = i\frac{d}{dt}U^\dagger \begin{pmatrix} \nu_{\rm e} \\ \nu_\mu \end{pmatrix}$$

$$= \begin{pmatrix} m_1^2/2E & 0 \\ 0 & m_2^2/2E \end{pmatrix} U^\dagger \begin{pmatrix} \nu_{\rm e} \\ \nu_\mu \end{pmatrix}.$$

Multiplying from the left by U and using unitarity, $UU^\dagger = 1$,

$$i\frac{d}{dt}\begin{pmatrix} \nu_{\rm e} \\ \nu_\mu \end{pmatrix} = U \begin{pmatrix} m_1^2/2E & 0 \\ 0 & m_2^2/2E \end{pmatrix} U^\dagger \begin{pmatrix} \nu_{\rm e} \\ \nu_\mu \end{pmatrix}.$$

As is clear by substitution, this equation has a solution

$$\begin{pmatrix} \nu_{\rm e}(t) \\ \nu_\mu(t) \end{pmatrix} = U \begin{pmatrix} e^{-i(m_1^2/2E)t} & 0 \\ 0 & e^{-i(m_2^2/2E)t} \end{pmatrix} U^\dagger \begin{pmatrix} \nu_{\rm e}(0) \\ \nu_\mu(0) \end{pmatrix}.$$

Since the masses m_1 and m_2 are presently unknown it is convenient to rewrite the equation of motion

$$i\frac{d}{dt}\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_\mu\end{pmatrix} = U\begin{pmatrix} m_1^2/2E & 0 \\ 0 & m_2^2/2E\end{pmatrix}U^\dagger\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_\mu\end{pmatrix}.$$

in terms of Δm^2, which is measurable.

- Adding a multiple of the unit matrix to the matrix will not modify observables (a trick that will be employed several times in what follows),
- so we may subtract $m_1^2/2E$ times the unit 2×2 matrix and use

$$\begin{pmatrix} 0 & 0 \\ 0 & \Delta m^2/2E\end{pmatrix} = \begin{pmatrix} m_1^2/2E & 0 \\ 0 & m_2^2/2E\end{pmatrix} - \begin{pmatrix} m_1^2/2E & 0 \\ 0 & m_1^2/2E\end{pmatrix}$$

to replace the equation of motion by the equivalent form

$$i\frac{d}{dt}\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_\mu\end{pmatrix} = U\begin{pmatrix} 0 & 0 \\ 0 & \Delta m^2/2E\end{pmatrix}U^\dagger\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_\mu\end{pmatrix},$$

where $\Delta m^2 \equiv m_2^2 - m_1^2$.

12.2.4 Propagation in Matter

The evolution equation

$$i\frac{d}{dt}\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_{\mu}\end{pmatrix}=U\begin{pmatrix}0 & 0\\ 0 & \Delta m^2/2E\end{pmatrix}U^{\dagger}\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_{\mu}\end{pmatrix},$$

is just a reformulation of our previous treatment of neutrinos propagating in vacuum.

- Let us now add a charged-current interaction with matter.
- By previous arguments the charged current couples
  - only elastically and
  - only to electron neutrinos.
- so we add to the evolution equation an interaction potential given by

  $$V(t)=\sqrt{2}G_{\mathrm{F}}n_{\mathrm{e}}(t),$$

  which modifies the equation of motion to

$$i\frac{d}{dt}\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_{\mu}\end{pmatrix}=\left[U\begin{pmatrix}0 & 0\\ 0 & \Delta m^2/2E\end{pmatrix}U^{\dagger}+\begin{pmatrix}V(t) & 0\\ 0 & 0\end{pmatrix}\right]\begin{pmatrix}\nu_{\mathrm{e}}\\ \nu_{\mu}\end{pmatrix}$$

As shown in a Problem, the equation of motion may be expressed as

$$\begin{aligned} i\frac{d}{dt}\begin{pmatrix} \nu_{\mathrm{e}} \\ \nu_{\mu} \end{pmatrix} &= \left[U \begin{pmatrix} 0 & 0 \\ 0 & \Delta m^2/2E \end{pmatrix} U^{\dagger} + \begin{pmatrix} V(t) & 0 \\ 0 & 0 \end{pmatrix} \right] \begin{pmatrix} \nu_{\mathrm{e}} \\ \nu_{\mu} \end{pmatrix} \\ &= \begin{pmatrix} V & \dfrac{\Delta m^2}{4E}\sin 2\theta \\ \dfrac{\Delta m^2}{4E}\sin 2\theta & \dfrac{\Delta m^2}{2E}\cos 2\theta \end{pmatrix} \begin{pmatrix} \nu_{\mathrm{e}} \\ \nu_{\mu} \end{pmatrix} \\ &\equiv M \begin{pmatrix} \nu_{\mathrm{e}} \\ \nu_{\mu} \end{pmatrix}. \end{aligned}$$

This is the required equation of motion in matter but it is conventional to write the *mass matrix*

$$M \equiv \begin{pmatrix} V & \dfrac{\Delta m^2}{4E}\sin 2\theta \\ \dfrac{\Delta m^2}{4E}\sin 2\theta & \dfrac{\Delta m^2}{2E}\cos 2\theta \end{pmatrix}$$

appearing in it in a more symmetric form by using that

> *A multiple of the unit matrix may be subtracted from M without affecting the values of quantum observables.*

First define

$$A \equiv 2EV = 2\sqrt{2}EG_{\mathrm{F}}n_{\mathrm{e}},$$

(which has units of mass squared) and then subtract

$$\frac{A}{4E} + \left(\frac{\Delta m^2}{4E}\right)\cos 2\theta$$

multiplied by the unit matrix (this has *no effect on observables!*) to give the mass matrix in traceless form

$$M = \frac{\pi}{L}\begin{pmatrix} \chi - \cos 2\theta & \sin 2\theta \\ \sin 2\theta & \cos 2\theta - \chi \end{pmatrix},$$

where the dimensionless *charged-current coupling strength* χ is defined by

$$\chi \equiv \frac{L}{\ell_{\mathrm{m}}} = \frac{2EV}{\Delta m^2} \qquad \ell_{\mathrm{m}} \equiv \frac{\sqrt{2}\pi}{G_{\mathrm{F}}n_{\mathrm{e}}} \qquad L \equiv \frac{4\pi E}{\Delta m^2},$$

with

- L the *vacuum oscillation length* and
- ℓ_{m} an additional contribution to the oscillation length caused by the matter interaction called the *refraction length*.

Electron neutrinos in the Sun *interact only through elastic forward scattering*.

- Thus the *effect of the medium* on ν_e propagation can be described as *a refraction*.
- The refraction is characterized by an *index of refraction*

  $$n_{\text{ref}} = 1 + \frac{V}{p},$$

  where

  $$V = \sqrt{2} G_F n_e \qquad p = |p|$$

- This is analogous to refraction of light in a medium, except that the ν index of refraction *depends on flavor*.
- The quantity ℓ_m is termed the *refraction length*.
- It is the distance over which an *additional phase of* 2π *is acquired through refraction* in the matter.

Notice that in vacuum

- $n_e \to 0$ so that $\ell_m \to \infty$ and
- the coupling term $\chi \equiv L/\ell_m$ vanishes,
- so we recover a mass matrix characteristic of vacuum oscillations.

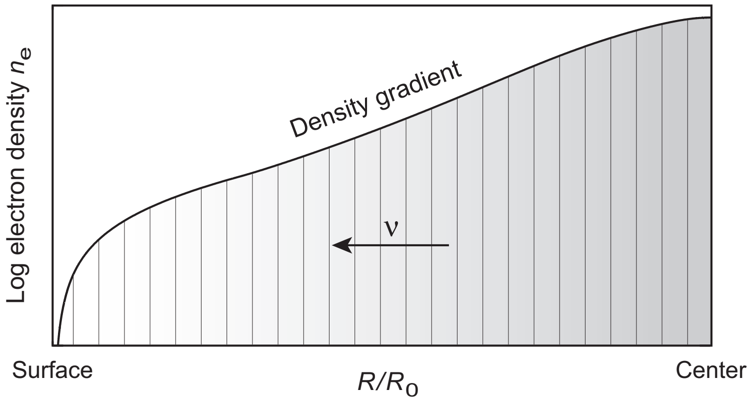


Figure 12.3: Solar density gradient. Neutrinos are produced near the center at high density and propagate out through regions of decreasing density. In a given concentric layer, the density may be assumed constant.

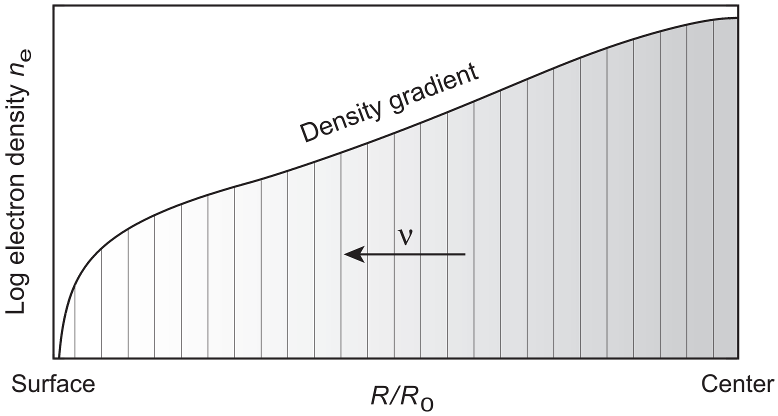
12.3 Solutions in Matter

For a fixed density the mass eigenstates in matter generally will

- differ from the mass eigenstates in vacuum because of V.
- They may be found by diagonalizing (finding the eigenvalues) of the mass matrix M at that density.
- However, the interaction V depends on the density.
- Thus in the Sun mass eigenstates in matter at one position will generally not be eigenstates at another position.

We may

- divide the Sun up into *concentric layers*, with
- density assumed to be *constant in a layer*, as illustrated in Fig. 12.3.



Our strategy will be to

- calculate the mass eigenstates within a single layer assuming it to have a constant density, and then
- consider how to determine the evolution of neutrino states as they propagate through successive layers of decreasing density on the way out of the Sun.

12.3.1 Mass Eigenvalues for Constant Density

At constant density, the problem resembles vacuum oscillations but with a different potential.

- The time-evolved mass states in matter, $|\nu_1^{\mathrm{m}}\rangle$ and $|\nu_2^{\mathrm{m}}\rangle$, may be obtained by
- diagonalizing the mass matrix at the current time.
- This gives two eigenvalues $\lambda_\pm$ at the current density,

$$\lambda_\pm = \left(\frac{m_1^2 + m_2^2}{2} + \frac{\Delta m^2}{2}\chi\right) \pm \frac{\Delta m^2}{2}\sqrt{(\cos 2\theta - \chi)^2 + \sin^2 2\theta}.$$

- The splitting between the two eigenstates is given by the second term.
- It reaches a minimum at the density where $\chi = \cos 2\theta$.
- As for the vacuum case, the mass eigenstates in matter, $|\nu_1^{\mathrm{m}}\rangle$ and $|\nu_2^{\mathrm{m}}\rangle$ at fixed time t, are assumed to be related to the flavor eigenstates by a unitary transformation

$$\begin{pmatrix} \nu_{\mathrm{e}} \\ \nu_\mu \end{pmatrix} = U_{\mathrm{m}}(t) \begin{pmatrix} \nu_1^{\mathrm{m}} \\ \nu_2^{\mathrm{m}} \end{pmatrix},$$

  where $U_{\mathrm{m}}(t)$ is a unitary matrix that

  - Differs from the vacuum transformation matrix U,
  - depends on time, and
  - is yet to be determined.

12.3.2 The Matter Mixing Angle θ_m

The matrix $U_\mathrm{m}(t)$

- depends on time and
- can be parameterized as for the vacuum mixing, but in terms of a time-dependent matter mixing angle $\theta_\mathrm{m}(t)$:

$$U_\mathrm{m} = \begin{pmatrix} \cos\theta_\mathrm{m} & \sin\theta_\mathrm{m} \\ -\sin\theta_\mathrm{m} & \cos\theta_\mathrm{m} \end{pmatrix} \qquad U_\mathrm{m}^\dagger = \begin{pmatrix} \cos\theta_\mathrm{m} & -\sin\theta_\mathrm{m} \\ \sin\theta_\mathrm{m} & \cos\theta_\mathrm{m} \end{pmatrix}.$$

The relationship of the matter mixing angle θ_m and the vacuum mixing angle θ at time t

- can be established by requiring that a similarity transform by $U_\mathrm{m}(t)$ diagonalize the mass matrix at that density,
- with the diagonal elements being the time-dependent eigenvalues in matter $E_1(t)$ and $E_2(t)$,

$$U_\mathrm{m}^\dagger(t) M U_\mathrm{m}(t) = \begin{pmatrix} E_1(t) & 0 \\ 0 & E_2(t) \end{pmatrix}.$$

Inserting explicit forms of U, $U^\dagger$, and M gives a matrix equation satisfied only if θ_m and θ are related by

$$\tan 2\theta_\mathrm{m} = \frac{\sin 2\theta}{\cos 2\theta \pm \chi} = \frac{\tan 2\theta}{1 \pm \chi/\cos 2\theta},$$

- where the plus sign is for $m_1 > m_2$ and
- the negative sign is for $m_1 < m_2$.

From the equation

$$\tan 2\theta_{\mathrm{m}} = \frac{\sin 2\theta}{\cos 2\theta \pm \chi} = \frac{\tan 2\theta}{1 \pm \chi / \cos 2\theta},$$

- in vacuum $\theta_{\mathrm{m}} = \theta$ because for vanishing electron density $\ell_{\mathrm{m}} \to \infty$ and $\chi = L/\ell_{\mathrm{m}} \to 0$, but
- in matter the mixing angle will be modified from its vacuum value by an amount that depends on density.

12.3.3 The Matter Oscillation Length L_m

The oscillation length in vacuum

$$L = \frac{4\pi E}{\Delta m^2}$$

- is proportional to the inverse of the mass-squared difference Δm^2 between the states participating in the oscillation.
- In matter the neutrino effective mass is altered by interaction with the medium and the vacuum mass-squared difference is rescaled,

$$\Delta m^2 \to f(\chi)\Delta m^2,$$

where from the splitting of the two eigenvalues in

$$\lambda_\pm = \left(\frac{m_1^2 + m_2^2}{2} + \frac{\Delta m^2}{2}\chi\right) \pm \frac{\Delta m^2}{2}\sqrt{(\cos 2\theta - \chi)^2 + \sin^2 2\theta}.$$

we deduce that

$$f(\chi) = \sqrt{(\cos 2\theta - \chi)^2 + \sin^2 2\theta} = \sqrt{1 - 2\chi\cos 2\theta + \chi^2}.$$

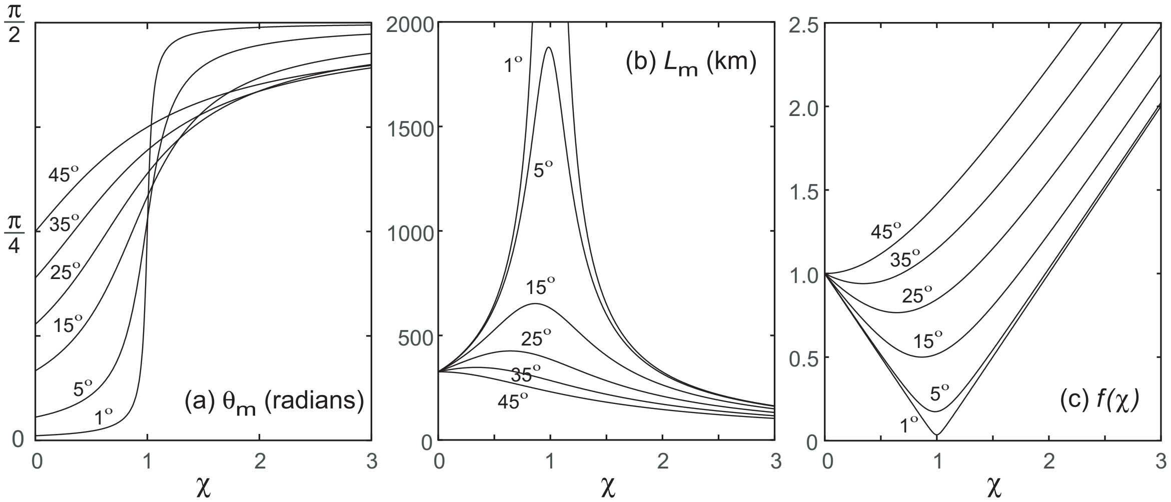


Figure 12.4: (a) Mixing angle in matter $\theta_{\rm m}(\chi)$, (b) oscillation length in matter $L_{\rm m}(\chi)$, and (c) scaling factor $f(\chi)$ as a function of the matter coupling χ. All calculations assumed $E = 10\,\text{MeV}$ and $\Delta m^2 = 7.6\times 10^{-5}\,\text{eV}^2$, and curves are marked with the assumed vacuum mixing angle θ.

Hence the *oscillation length in matter* $L_{\rm m}$ is given by

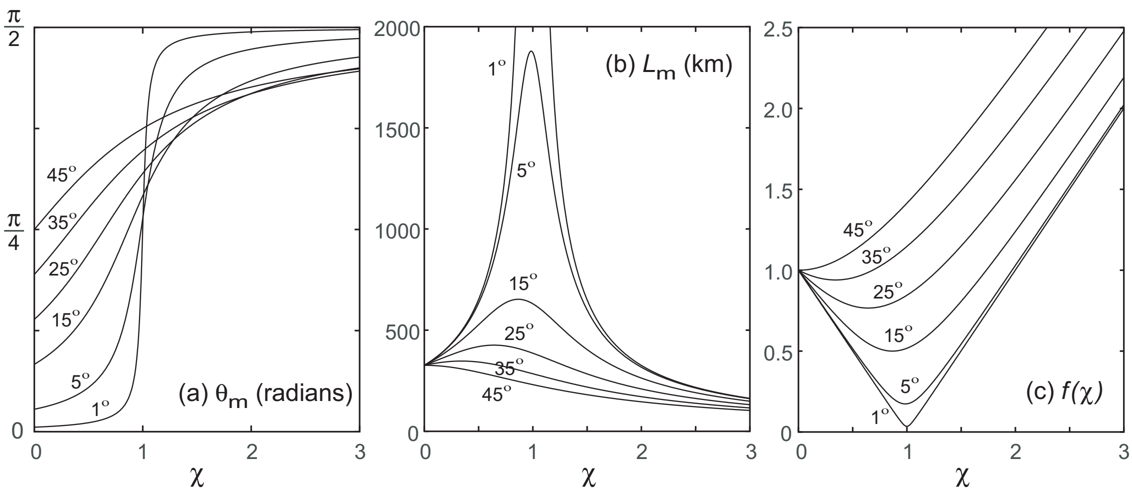
$$L_{\rm m} = \frac{4\pi E}{f(\chi)\Delta m^2} = \frac{L}{f(\chi)} = \frac{L}{\sqrt{(\cos 2\theta - \chi)^2 + \sin^2 2\theta}},$$

where

$$f(\chi) = \sqrt{(\cos 2\theta - \chi)^2 + \sin^2 2\theta},$$

which reduces to the vacuum oscillation length L if the interaction χ vanishes.

The variations of $\theta_{\rm m}$, $L_{\rm m}$, and f with the dimensionless coupling χ are illustrated in Fig. 12.4 for several values of the vacuum mixing angle θ.



From (a) in the above figure,

- the matter mixing angle $\theta_{\rm m}$ reduces to the vacuum mixing angle θ for vanishing coupling, but
- $\theta_{\rm m} \to \frac{\pi}{2}$ at large coupling for *any* θ.

From (b) in the figure above

- the matter oscillation length is equal to the vacuum oscillation length at zero coupling, but
- increases to a maximum at the coupling strength where $\theta_{\rm m} = \frac{\pi}{4}$, and then decreases again.

The coupling strength at which $L_{\rm m}$ is maximal

- coincides with highest rate of change in $\theta_{\rm m}$, suggesting
- something special about the density where $\theta_{\rm m} = \frac{\pi}{4}$.

We will address the implications of this observation shortly.

12.3.4 Flavor Conversion in Constant-Density Matter

In matter of *constant density*

- the electron neutrino state after a time t becomes

$$
\begin{aligned}
|\nu(t)\rangle = (\cos^2\theta_{\rm m} e^{-iE_1t} + \sin^2\theta_{\rm m} e^{-iE_2t})\,|\nu_{\rm e}\rangle \\
+ \sin\theta_{\rm m}\cos\theta_{\rm m}(-e^{-iE_1t} + e^{-iE_2t})\,|\nu_\mu\rangle\,,
\end{aligned}
$$

- This is analogous to the corresponding vacuum equations but with the vacuum mixing angle θ replaced by the matter mixing angle $\theta_{\rm m}$.
- Hence for a constant density $n_{\rm e}$ the flavor conservation and flavor retention probabilities are given by the corresponding vacuum equations with the replacements $\theta \rightarrow \theta_{\rm m}$ and $L \rightarrow L_{\rm m}$,

$$
\begin{aligned}
P(\nu_{\rm e}\to\nu_{\rm e}, r) &= 1 - \sin^2 2\theta_{\rm m}\sin^2\left(\frac{\pi r}{L_{\rm m}}\right)\,, \\
P(\nu_{\rm e}\to\nu_\mu, r) &= 1 - P(\nu_{\rm e}\to\nu_{\rm e}, r)
\end{aligned}
$$

- The corresponding classical averages are

$$
\bar P(\nu_{\rm e}\to\nu_{\rm e}) = 1 - \tfrac{1}{2}\sin^2 2\theta_{\rm m} \qquad \bar P(\nu_{\rm e}\to\nu_\mu) = \tfrac{1}{2}\sin^2 2\theta_{\rm m},
$$

  which are valid when the uncertainty in distance between source and detection exceeds the oscillation length.

12.4 The MSW Resonance Condition

From

$$P(\nu_{\rm e} \to \nu_{\rm e}, r) = 1 - \sin^2 2\theta_{\rm m} \sin^2\left(\frac{\pi r}{L_{\rm m}}\right) ,$$

optimal flavor mixing occurs

- whenever $\sin^2 2\theta_{\rm m}$ achieves its maximum value of unity,
- which occurs when $|\theta_{\rm m}| = \frac{\pi}{4}$.

The most significant property of

$$\tan 2\theta_{\rm m} = \frac{\sin 2\theta}{\cos 2\theta \pm \chi} = \frac{\tan 2\theta}{1 \pm \chi / \cos 2\theta},$$

is that

- if Δm^2 and L are positive (which requires that $m_1 < m_2$),
- then $\tan 2\theta_{\rm m} \to \pm\infty$ and $\theta_{\rm m} \to \frac{\pi}{4}$ whenever the coupling strength satisfies

  $$\chi = \frac{L}{\ell_{\rm m}} = \cos 2\theta,$$

  which occurs when the electron density satisfies

  $$n_{\rm e} = \frac{\cos 2\theta \Delta m^2}{2\sqrt{2} G_{\rm F} E} \equiv n_{\rm e}^{\rm R}.$$

This is a *resonance condition*.

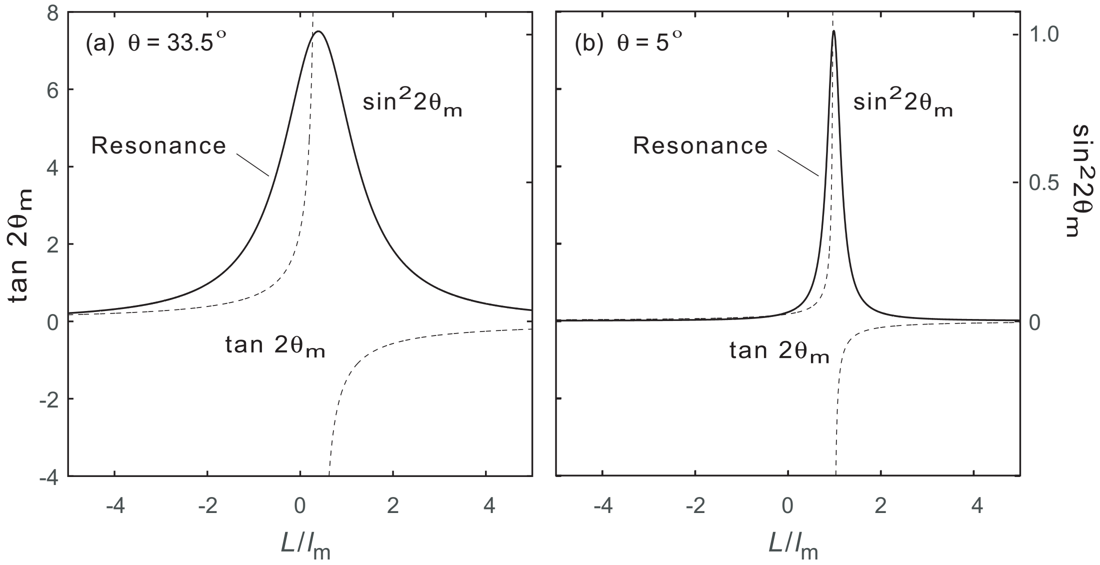


Figure 12.5: The MSW resonance condition for two values of the vacuum mixing angle θ. When $L/\ell_{\rm m} \to \cos 2\theta$ the denominator of Eq. (12.4) goes to zero, $\tan 2\theta_{\rm m}$ goes to $\pm\infty$ so that $|\theta_{\rm m}| \to \frac{\pi}{4}$, and the flavor conversion probability $\sin^2 2\theta_{\rm m}$ attains its maximum value. Thus, at the resonance Eq. (12.4) indicates that large flavor conversion can be obtained for any non-vanishing vacuum oscillation angle θ.

The resonance condition

$$\chi = \frac{L}{\ell_{\rm m}} = \cos 2\theta \qquad \tan 2\theta_{\rm m} \to \pm\infty \qquad \theta_{\rm m} \to \frac{\pi}{4}$$

is shown in Fig. 12.5. It leads to maximal mixing between electron neutrinos and muon neutrinos, with a $\nu_{\rm e}$ survival

$$P(\nu_{\rm e} \to \nu_{\rm e}, r) = 1 - \sin^2\left(\frac{\pi r}{L_{\rm m}}\right) \qquad \text{(at resonance)},$$

and an oscillation length at resonance $L_{\rm m}^{\rm R}$ given by

$$L_{\rm m}^{\rm R} = L_{\rm m}(\chi = \cos 2\theta) = \frac{L}{\sin 2\theta}.$$

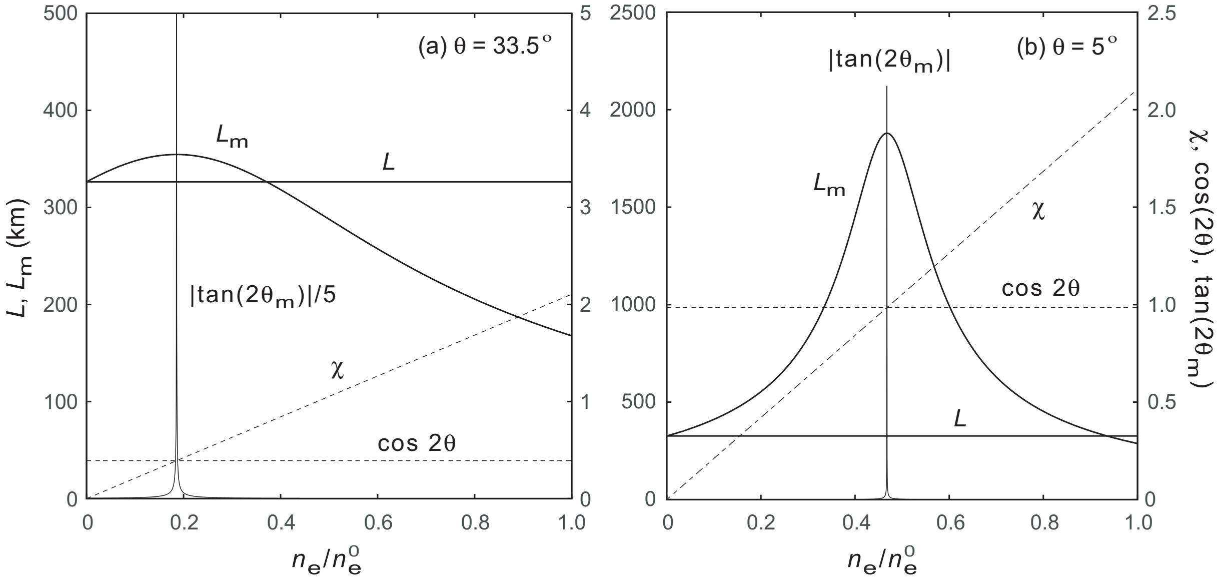


Figure 12.6: Resonance parameters versus the electron number density n_e in units of the central solar value $n_\mathrm{e}^0 \sim 6.3 \times 10^{25}\,\mathrm{cm}^{-3}$ for $\theta = 33.5^\circ$ and 5°, with $\Delta m^2 = 7.6 \times 10^{-5}\,\mathrm{eV}^2$ and $E = 10\,\mathrm{MeV}$. The coupling strength $\chi = L/\ell_\mathrm{m}$ is linear in the density. Intersection of the dashed lines specifies the electron density giving the resonance condition.

The condition

$$\chi = \frac{L}{\ell_\mathrm{m}} = \cos 2\theta \qquad \tan 2\theta_\mathrm{m} \to \pm\infty \qquad \theta_\mathrm{m} \to \frac{\pi}{4}$$

defines the *Mikheyev–Smirnov–Wolfenstein* or *MSW resonance*.

- *No matter how small* the vacuum mixing angle θ, if it is not zero there is some critical value n_e^R of the electron density where the resonance condition is satisfied and
- at the resonance, *maximal flavor mixing* ensues.

The important resonance parameters are plotted in Fig. 12.6 as a function of electron density for two values of θ.

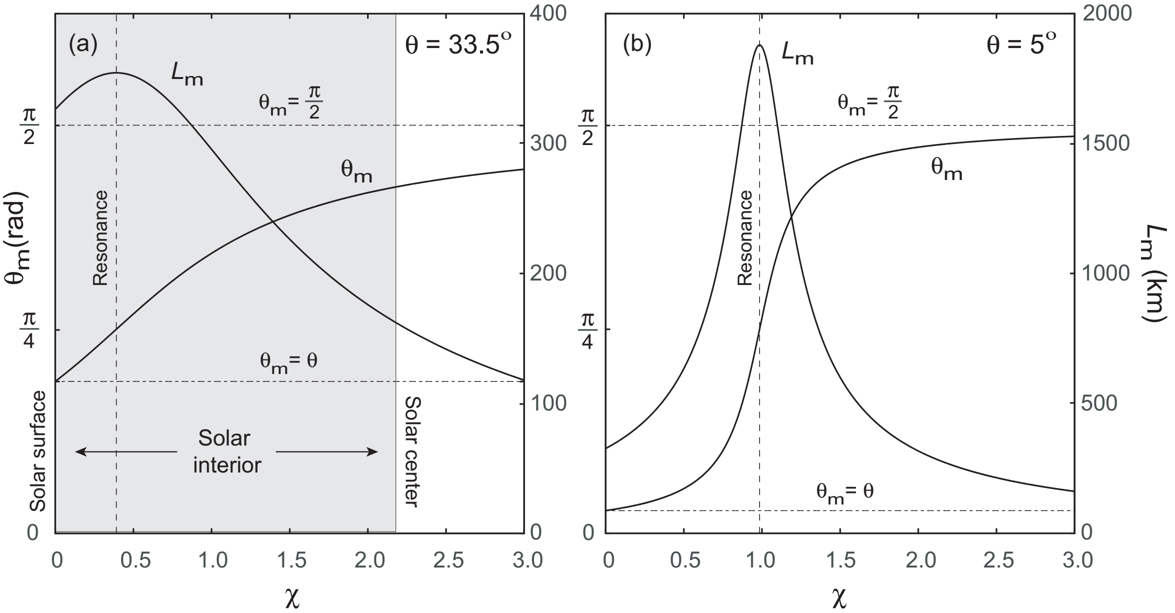


Figure 12.7: The matter mixing angle $\theta_{\rm m}$ as a function of the dimensionless coupling strength $\chi \equiv L/\ell_{\rm m}$ for vacuum mixing angles of (a) $\theta = 33.5^\circ$ and (b) $\theta = 5^\circ$. Also shown is the oscillation length in matter $L_{\rm m}$, which has a maximum at the position of the MSW resonance (see §12.4), marked by the dashed vertical line. The oscillation length was computed assuming $E = 10\,\text{MeV}$ and $\Delta m^2 = 7.6 \times 10^{-5}\,\text{eV}^2$. Case (a) is realistic for solar neutrinos and the density at the center of the Sun corresponds to $\chi \sim 2.13$. Hence the shaded region on the left side of (a) indicates the range of coupling strengths available to electron neutrinos in the interior of the Sun.

The effect of the MSW resonance on variation of the matter mixing angle $\theta_{\rm m}$ and the oscillation length in matter $L_{\rm m}$ are illustrated for a small and large angle solution in Fig. 12.7.

- The values of $\theta_{\rm m}$ and $L_{\rm m}$ will vary with the solar depth since they depend on the number density $n_{\rm e}$ through χ.
- $\theta_{\rm m} \to \theta$ as the electron density tends to zero, while
- in the opposite limit of very large electron density $\theta_{\rm m} \to \frac{\pi}{2}$.

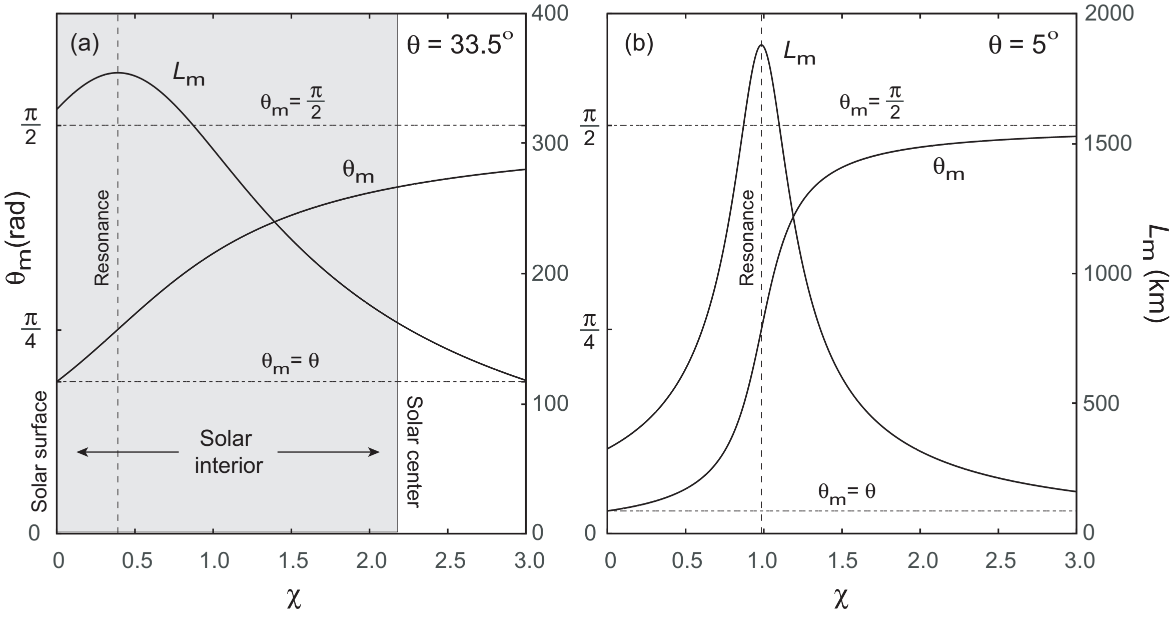


Figure (a) above corresponds to parameters valid for solar neutrinos.

- At the solar center ($\chi \sim 2.13$) the matter mixing angle is $\theta_{\rm m} \sim 76°$,
- compared with a vacuum mixing angle $33.5°$ at the solar surface.

Conversely, for Figure (b) above

- the matter mixing angle at a density corresponding to the solar center is $\theta \sim 86°$,
- with $\theta = 5°$ at the solar surface.

We have assumed $m_1 < m_2$ in deriving the MSW resonance. If instead $m_1 > m_2$

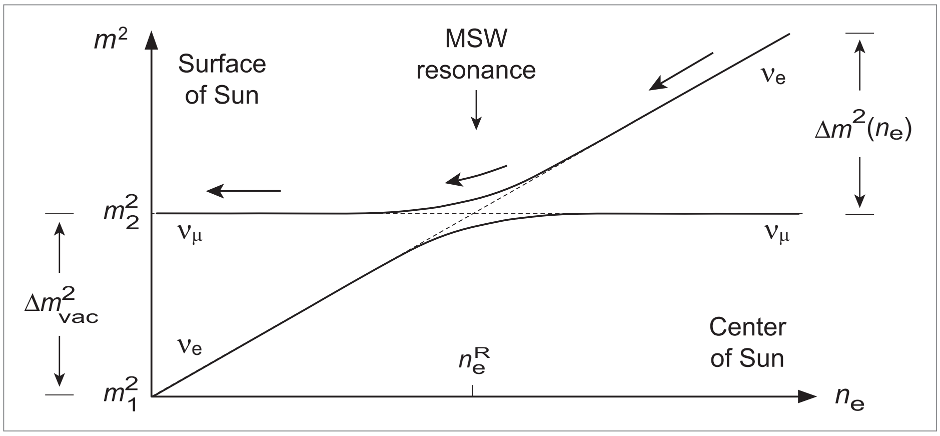
- there is *no resonance* for ν_e.
- In that case there is a resonance instead for the *electron antineutrino* $\bar{\nu}_e$.
- The Sun emits primarily neutrinos and not antineutrinos.
- Thus a discussion of antineutrinos will be omitted in the present context.

However, antineutrino oscillations could occur in core collapse supernovae or neutron star mergers, where all flavors of ν and $\bar{\nu}$ are produced in abundance.

12.5 Resonant Flavor Conversion

- If, for example,
  - $m_1 < m_2$ and θ is small so that $|\nu_\mathrm{e}\rangle \simeq |\nu_1^\mathrm{m}\rangle$, and
  - the electron density in the central part of the Sun where the neutrino is produced satisfies $n_\mathrm{e} > n_\mathrm{e}^\mathrm{R}$,
- a neutrino leaving the Sun will inevitably encounter the MSW resonance while on its way out of the Sun.
- If the change in density is sufficiently slow (the *adiabatic condition* discussed below),
- the ν_e flux produced in the core can be almost entirely converted to ν_μ by the MSW resonance near the radius where the resonance condition is satisfied.

The MSW resonance conversion of flavors can be viewed as an *adiabatic level crossing*



- If the level crossing is adiabatic,
  - a neutrino that *starts out as a* ν_e near the center (high density on the right side of the figure)
  - *changes adiabatically into a* ν_μ by the time it exits the Sun (low density in the left side of the figure).
- That is, the neutrino *follows the upper curved trajectory though the resonance* in the level-crossing region, as indicated by the arrows.

Therefore, the neutrino can *emerge from the Sun in a completely different flavor state* than the one in which it was created.

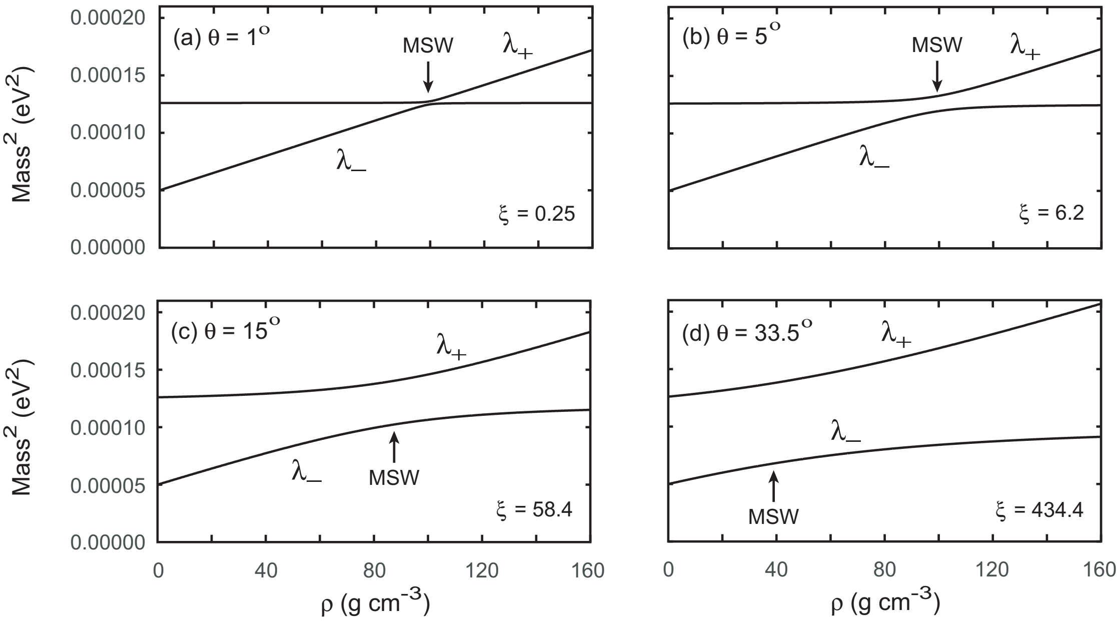


Figure 12.8: Solutions $\lambda_\pm$ of the MSW eigenvalue problem as a function of mass density. Each case corresponds to the choices $\Delta m^2 = 7.6 \times 10^{-5}\,\mathrm{eV}^2$ and $E = 10\,\mathrm{MeV}$, but to different values of the vacuum mixing angle θ. The individual neutrino masses are presently unknown but for purposes of illustration $m_1^2 = 5 \times 10^{-5}\,\mathrm{eV}^2$ has been assumed in vacuum, so that $m_2^2 = m_1^2 + \Delta m^2 = 1.26 \times 10^{-4}\,\mathrm{eV}^2$. The critical density leading to the MSW resonance (corresponding to minimum splitting between the eigenvalues) and the value of the adiabaticity parameter $\xi = \delta r_\mathrm{R}/L_\mathrm{m}^\mathrm{R}$ are indicated for each case. Realistic conditions in the Sun are expected to imply the very adiabatic crossing exhibited in case (d).

Figure 12.8 illustrates solutions of the MSW eigenvalue problem for different choices of θ as a function of mass density.

- Small θ implies sharp level crossings and larger θ implies adiabatic (strongly-avoided) crossings.
- The strongly-avoided level crossing in Fig. (d) above is expected to apply for the Sun.

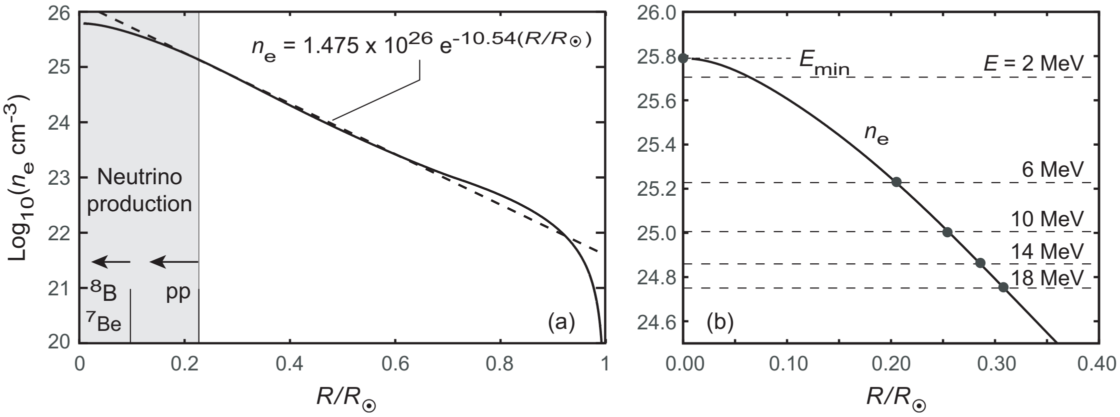
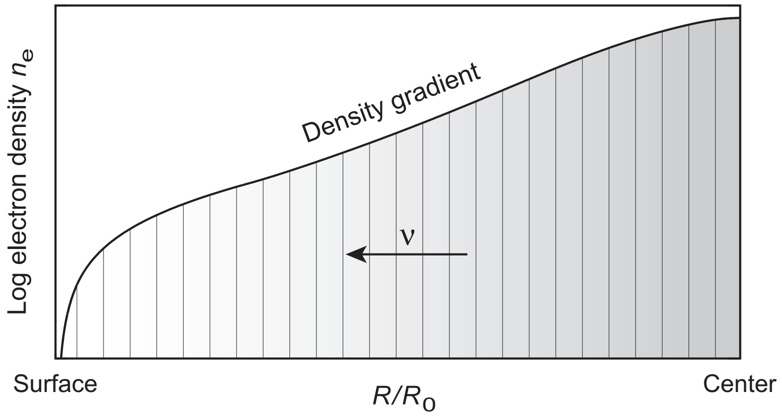


Figure 12.9: (a) Electron number density as a function of fractional solar radius from the Standard Solar Model. The dashed line is an exponential approximation that will be employed in discussing the MSW effect. Regions of primary neutrino production in the PP chains are indicated. (b) Radius where the MSW critical density for a 2-flavor model is realized (dots at intersection of dashed lines with the curve for n_e) for neutrinos of energies ranging from 2 to 18 MeV. A vacuum mixing angle $\theta = 35^\circ$ and $\Delta m^2 c^4 = 7.5 \times 10^{-5}\,\mathrm{eV}^2$ have been assumed. The minimum energy of an electron neutrino $E_{\mathrm{min}} \sim 1.6\,\mathrm{MeV}$ that could be produced in the Sun and still encounter the MSW resonance is indicated.

The number density of electrons in the Sun is illustrated in Fig. 12.9(a), along with an exponential approximation.

- In Fig. 12.9(b) the locations where electron neutrinos of various energies would encounter the MSW resonance condition are illustrated.
- Only neutrinos with energy larger than some minimum energy $E_{\mathrm{min}} \sim 1.6\,\mathrm{MeV}$ can experience the resonance.
- Thus the MSW effect should be more efficient at converting higher-energy neutrinos.
- We shall see that this agrees with data.



12.6 Propagation in Matter of Varying Density

We are ready to consider realistic solar neutrino propagation.

- A neutrino produced in the center will encounter decreasing density as it travels toward the solar surface, as illustrated in the figure above.
- The neutrino flavor evolution will be governed by the analog of the differential equations for vacuum propagation,
- but with $U \to U_{\rm m}(t)$ since the flavor–mass basis transformation now depends on time.

$$i\frac{d}{dt}\left[U_{\rm m}(t)\begin{pmatrix}\nu_1(t)\\ \nu_2(t)\end{pmatrix}\right] = \frac{1}{2E}U_{\rm m}(t)\begin{pmatrix}m_1^2 & 0\\ 0 & m_2^2\end{pmatrix}\begin{pmatrix}\nu_1(t)\\ \nu_2(t)\end{pmatrix},$$

where both the wavefunctions and the transformation matrix $U_{\rm m}$ are indicated explicitly to depend on the time.

Taking the derivative of the product in brackets on the left side in the equation

$$i\frac{d}{dt}\left[U_{\mathrm{m}}(t)\begin{pmatrix}\nu_1(t)\\ \nu_2(t)\end{pmatrix}\right]=\frac{1}{2E}U_{\mathrm{m}}(t)\begin{pmatrix}m_1^2 & 0\\ 0 & m_2^2\end{pmatrix}\begin{pmatrix}\nu_1(t)\\ \nu_2(t)\end{pmatrix},$$

and multiplying the equation from the left by $U_{\mathrm{m}}^{\dagger}$ gives

$$i\frac{d}{dt}\begin{pmatrix}\nu_1\\ \nu_2\end{pmatrix}=\begin{pmatrix}-\Delta m^2/4E & -i\,\dot{\theta}_{\mathrm{m}}\\ i\,\dot{\theta}_{\mathrm{m}} & \Delta m^2/4E\end{pmatrix}\begin{pmatrix}\nu_1\\ \nu_2\end{pmatrix},$$

- where $\dot{\theta}_{\mathrm{m}}\equiv\dfrac{d\theta_{\mathrm{m}}}{dt}$, and
- the constant
$$\frac{(m_1^2+m_2^2)}{4E}$$
times the unit matrix has been subtracted from the matrix on the right side.
- (This subtraction does not affect observables).

The earlier statement that mass eigenstates at some density will not be eigenstates at another density may now be quantified.

- If the mass matrix

$$M = \begin{pmatrix} -\Delta m^2/4E & -i\,\dot{\theta}_{\rm m} \\ i\,\dot{\theta}_{\rm m} & \Delta m^2/4E \end{pmatrix}$$

  in the equation of motion

$$i\frac{d}{dt}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} -\Delta m^2/4E & -i\,\dot{\theta}_{\rm m} \\ i\,\dot{\theta}_{\rm m} & \Delta m^2/4E \end{pmatrix}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix},$$

  were diagonal,

- the neutrino would remain in its original mass eigenstate.
- Thus it is the off-diagonal terms $\sim \dot{\theta}_{\rm m} = d\theta_{\rm m}/dt$ that alter the mass eigenstates as the neutrino propagates.
- Generally the equation of motion must be solved numerically.
- However, if the off-diagonal terms are small relative to the diagonal terms, the mass matrix M may be approximated

$$M = \begin{pmatrix} -\Delta m^2/4E & -i\,\dot{\theta}_{\rm m} \\ i\,\dot{\theta}_{\rm m} & \Delta m^2/4E \end{pmatrix} \simeq \begin{pmatrix} -\Delta m^2/4E & 0 \\ 0 & \Delta m^2/4E \end{pmatrix},$$

- This is called the *adiabatic approximation*.

The adiabatic approximation

$$i\frac{d}{dt}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \simeq \begin{pmatrix} -\Delta m^2/4E & 0 \\ 0 & \Delta m^2/4E \end{pmatrix}\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix},$$

affords an analytical solution for neutrino flavor conversion in the Sun.

- It corresponds physically to the assumption that the matter mixing angle θ_{m} changes only slowly over a characteristic time for motion of the neutrino.
- A neutrino travels at nearly the speed of light.
- Therefore, $r \sim ct$ and the adiabatic condition also may be interpreted as a limit on the spatial gradient of θ_{m}.

These observations may be used to quantify the conditions appropriate for the adiabatic approximation.

12.7 The Adiabatic Criterion

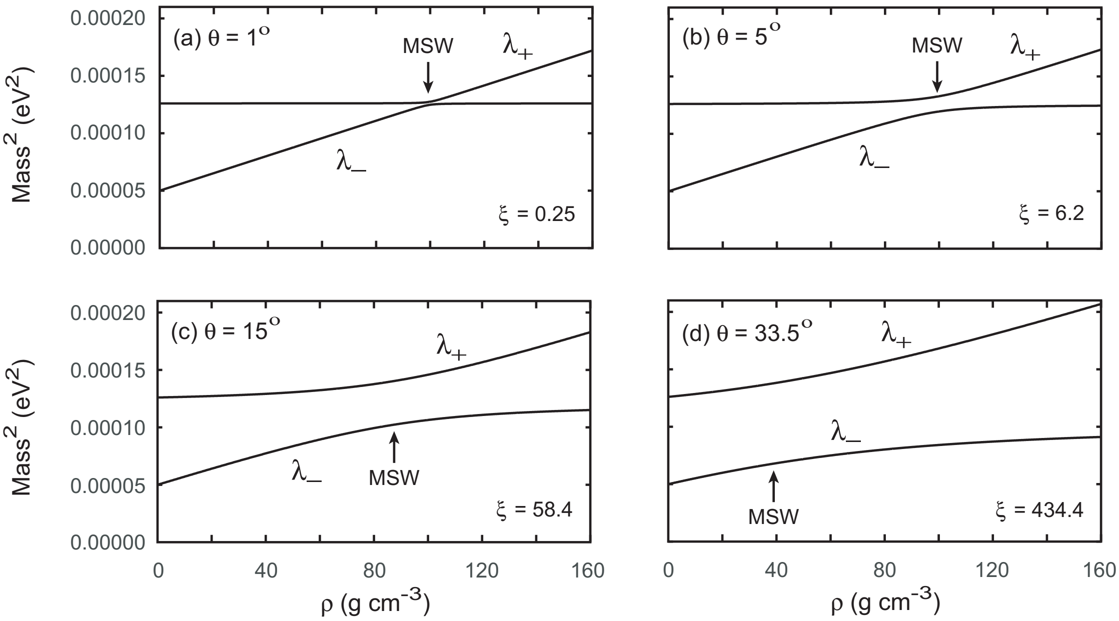
The adiabatic condition for resonant flavor conversion can be expressed as a requirement that

- the spatial width of the resonance layer δr_R (defined by the radial distance over which the resonance condition is approximately satisfied)
- be much greater than the oscillation wavelength in matter evaluated at the resonance, L_m^R.
- This can be characterized by an *adiabaticity parameter* ξ defined by

  $$\xi \equiv \frac{\delta r_\mathrm{R}}{L_\mathrm{m}^\mathrm{R}} \qquad \delta r_\mathrm{R} = \frac{n_\mathrm{e}^\mathrm{R}}{(dn_\mathrm{e}/dr)_\mathrm{R}} \tan 2\theta \qquad L_\mathrm{m}^\mathrm{R} = \frac{L}{\sin 2\theta},$$

  where

  - the label R denotes quantities evaluated at the resonance,
  - L is the vacuum oscillation length, and
  - θ is the vacuum oscillation angle.
- The adiabatic condition corresponds to requiring that $\xi \gg 1$.
- This implyies physically that if many oscillation lengths (in matter) fit within the resonance layer the adiabatic approximation is valid.



Values of ξ computed from

$$\xi \equiv \frac{\delta r_{\rm R}}{L_{\rm m}^{\rm R}} \qquad \delta r_{\rm R} = \frac{n_{\rm e}^{\rm R}}{(dn_{\rm e}/dr)_{\rm R}} \tan 2\theta \qquad L_{\rm m}^{\rm R} = \frac{L}{\sin 2\theta},$$

are indicated in the figure above.

- *Sharp level crossings* as in (a) are *non-adiabatic*, while
- *avoided level crossings* as in (d) are *highly adiabatic*.
- The MSW resonance can occur approximately adiabatically, even for relatively small θ [Example: case (b)].
- The actual Sun corresponds to (d), for which $\delta r_{\rm R} \gg L_{\rm m}^{\rm R}$.

The MSW resonance is expected to be *encountered adiabatically in the Sun*, optimizing the chance of resonant flavor conversion.

12.8 MSW Neutrino Flavor Conversion

The MSW resonance has been likened to the interaction of two tuning forks, one of which has a variable length.

- Suppose a vibrating tuning fork with variable length to be
  - brought close to another tuning fork and then
  - the length of the first is varied to match the second.
- Then the vibration of the first fork can be almost completely resonantly transferred to the second.
- However, this occurs only if the frequency of the variable fork is *changed slowly enough* (adiabatic condition).

Likewise, because the neutrino–matter interaction affects only the electron neutrino and it depends on density,

- a $\nu_{\rm e}$ matter wave in the Sun has a variable frequency while
- a ν_μ matter wave has roughly a constant frequency.
- As the $\nu_{\rm e}$ frequency varies with density it can become equal to that of the ν_μ at some depth in the Sun.
- This will lead to a resonance between the two matter waves and
- resonant flavor conversion at that depth.

As for tuning forks, large flavor conversion can occur *only if the adiabatic condition is satisfied*.

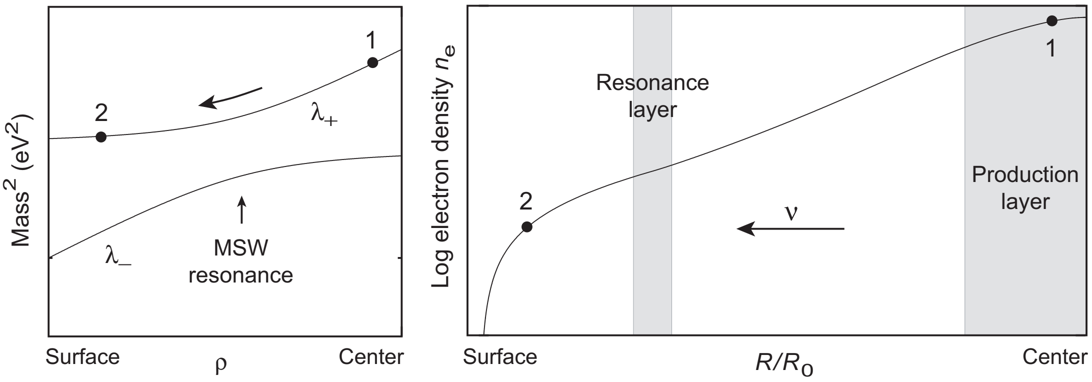
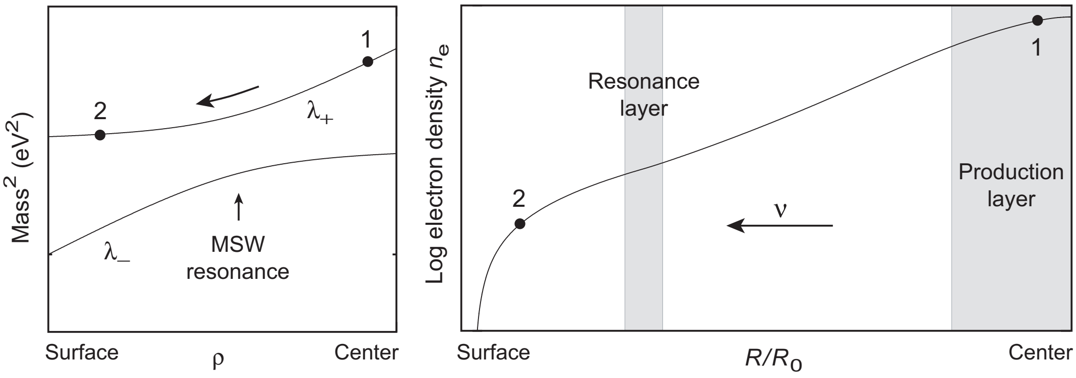


Figure 12.10: Schematic illustration of adiabatic flavor conversion by the MSW mechanism in the Sun. An electron neutrino is produced at Point 1, where the density lies above that of the MSW resonance, and propagates radially outward to Point 2, where the density lies below that of the resonance. The width of the resonance layer is assumed to be much larger than the matter oscillation length in the resonance layer, justifying the adiabatic approximation of Eq. (12.6). The widths of resonance and production layers are not meant to be to scale in this diagram.

12.8.1 Flavor Conversion in Adiabatic Approximation

Adiabatic flavor conversion is illustrated in Fig. 12.10.

- An electron neutrino is
  - produced at Point 1 near the center of the Sun and
  - propagates radially outward to Point 2.
- Detection is assumed to average over many oscillation lengths.
- Thus interference terms are washed out and our concern is with the classical (time-averaged) probability.



The *classical probability* in *adiabatic approximation* to be detected at Point 2 in the $|\nu_e\rangle$ flavor eigenstate is then given by

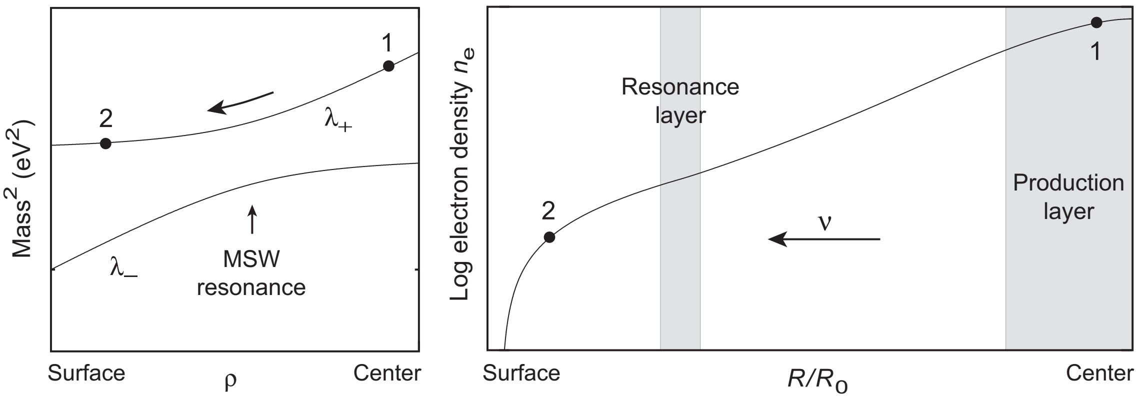
- generalizing the earlier result in vacuum (see Ch. 11)

$$\bar{P}(\nu_e \to \nu_e) = (1 \quad 0) \begin{pmatrix} \cos^2\theta(2) & \sin^2\theta(2) \\ \sin^2\theta(2) & \cos^2\theta(2) \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos^2\theta(1) & \sin^2\theta(1) \\ \sin^2\theta(1) & \cos^2\theta(1) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

  to the *classical adiabatic result in matter*

$$\bar{P}(\nu_e \to \nu_e) = (1 \quad 0) \begin{pmatrix} \cos^2\theta_m(2) & \sin^2\theta_m(2) \\ \sin^2\theta_m(2) & \cos^2\theta_m(2) \end{pmatrix} \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos^2\theta_m(1) & \sin^2\theta_m(1) \\ \sin^2\theta_m(1) & \cos^2\theta_m(1) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

- where $\theta_m(i) \equiv \theta_m(t_i)$ and
- $(1 \ \ 0)$ and $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ denote pure ν_e flavor states.



Evaluating the matrix products and using standard trigonometric identities gives for the *probability to remain a* ν_e,

$$\bar{P}(\nu_e \to \nu_e) = \tfrac{1}{2}\left[1 + \cos 2\theta_m(t_1) \cos 2\theta_m(t_2)\right].$$

- This result is valid (if the adiabatic condition is satisfied) for Point 2 anywhere outside Point 1, but
- in the specific case that *Point 2 lies at the solar surface* $\theta_m(t_2) \to \theta$ and
- the classical probability to detect the neutrino as an electron neutrino when it *exits the Sun* is

$$\bar{P}(\nu_e \to \nu_e) = \tfrac{1}{2}\left(1 + \cos 2\theta \cos 2\theta_m^0\right) \qquad \text{(at solar surface)},$$

- where θ is the *vacuum mixing angle* and
- $\theta_m^0 \equiv \theta_m(t_1)$ is the matter mixing angle *at the point of neutrino production*.

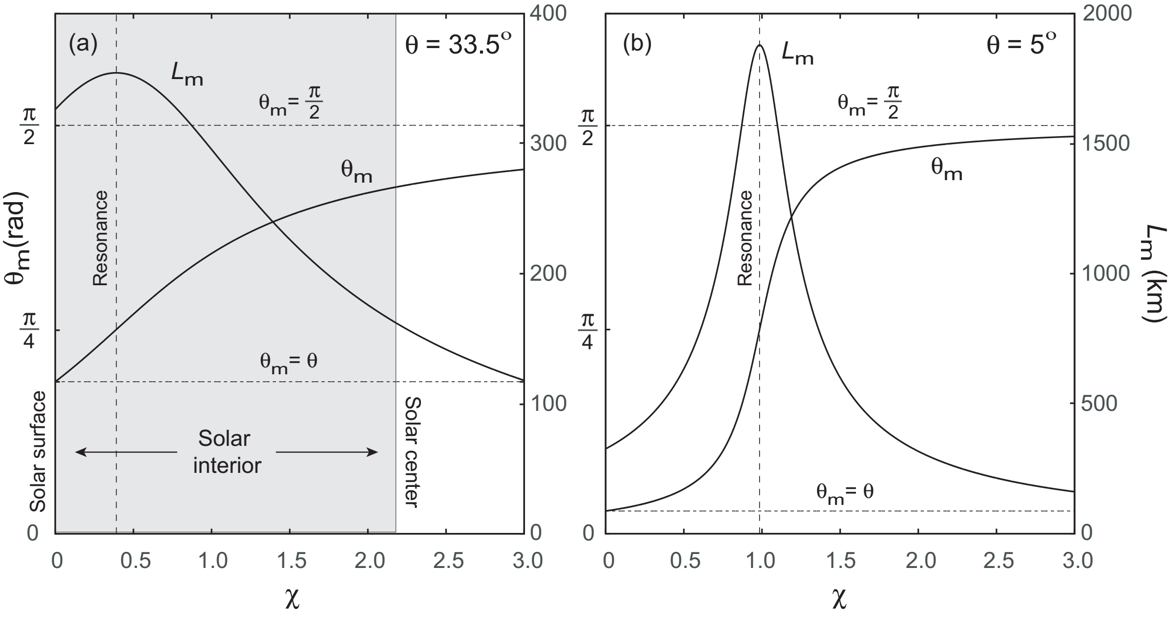
12.8.2 Adiabatic Conversion and the Mixing Angle

The remarkably concise result

$$\bar{P}(\nu_{\rm e} \to \nu_{\rm e}) = \tfrac{1}{2}(1 + \cos 2\theta \cos 2\theta_{\rm m}^{0}) \qquad \text{(at solar surface)},$$

has a simple physical interpretation.

- Because of the adiabatic assumption the *mass matrix is diagonal* for a neutrino propagating down the solar density gradient.
- Thus a neutrino produced in the λ_+ eigenstate *remains in that eigenstate* until it reaches the solar surface,
- with the flavor conversion resulting *only from the change of mixing angle* between production point and surface.
- Thus, the classical adiabatic probability $\bar{P}(\nu_{\rm e} \to \nu_{\rm e})$
  - is independent of the details of neutrino propagation and
  - *depends only on the mixing angles* at the point of production and point of detection.



Example: From Fig. (a) above left,

- At the Sun's center the *matter mixing angle* is $\theta_m^0 \sim 76°$.
- Thus an electron neutrino produced at the center of the Sun has a probability to be a ν_e when it exits the Sun of

$$\begin{aligned}\bar{P}(\nu_e \to \nu_e) &= \tfrac{1}{2}(1 + \cos 2\theta \cos 2\theta_m^0)\\ &= \tfrac{1}{2}[1 + \cos(2 \times 33.5°)\cos(2 \times 76°)]\\ &= 0.33,\end{aligned}$$

and a probability to be a ν_μ of

$$\bar{P}(\nu_e \to \nu_\mu) = 1 - \bar{P}(\nu_e \to \nu_e) = 0.67$$

Because of MSW, only $\sim \frac{1}{3}$ of 10-MeV solar neutrinos will still be ν_e when they exit the Sun.

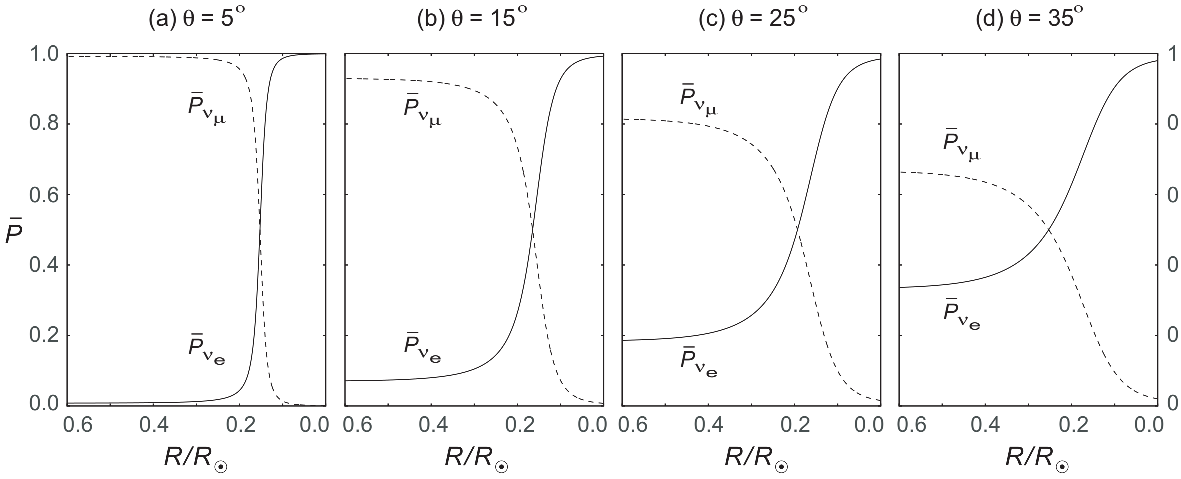


Figure 12.11: MSW flavor conversion vs. fraction of solar radius for four values of the vacuum mixing angle θ. Calculations are classical averages over local oscillations in adiabatic approximation using Eq. (12.8.2), assuming $\Delta m^2 = 7.6 \times 10^{-5}\,\text{eV}^2$ and $E = 10\,\text{MeV}$. Neutrinos were assumed to be produced in a ν_e flavor state at the center of the Sun (right side of diagram at $R/R_\odot = 0$). Solid curves show the classical electron-neutrino probability and dashed curves show the corresponding classical muon-neutrino probability.

Flavor conversion by the MSW mechanism for a 2-flavor model in adiabatic approximation is illustrated for four different values of the vacuum mixing angle θ in Fig. 12.11.

- The MSW resonance occurs at the radius corresponding to the intersection of the solid and dashed curves.
- Figure 12.11(d) approximates the situation expected for the Sun.

Table 12.1: Energy dependence of solar ν flavor conversion for $\theta = 35^\circ$

E (MeV)	14	10	6	2	1	0.70
P_{ν_e} (surface)	0.33	0.33	0.34	0.40	0.47	0.50
$R^{\mathrm{R}}/R_\odot$	0.28	0.25	0.20	0.10	0.03	0.0

The table above gives the *dependence on neutrino energy of flavor conversion in the Sun*, assuming a vacuum mixing angle of $\theta = 35^\circ$.

12.9 Resolution of the Solar Neutrino Problem

A series of experiments have *resolved the solar neutrino problem*.

- These experiments demonstrate rather directly that *neutrinos undergo flavor oscillations*.
- This, in turn then implies that at least some *neutrinos have mass*.
- Detailed comparison of these experiments indicates that *solar electron neutrinos are being converted to muon neutrinos* by neutrino oscillations,
- If *all flavors of neutrinos* coming from the Sun are detected the solar neutrino deficit relative to the Standard Solar Model disappears.
- The favored oscillation scenario is
  - *MSW resonance conversion* in the Sun,
  - but for a *large vacuum mixing angle solution.*

Let's now describe briefly the experiments that have led to these rather remarkable conclusions.

12.9.1 Super Kamiokande Observation of Flavor Oscillation

The Super Kamiokande detector has been used to observe neutrinos produced in atmospheric cosmic ray showers.

- When high-energy cosmic rays hit the atmosphere, they generate showers of mesons that decay to muons, electrons, positrons, and neutrinos.
- Theory assuming no physics beyond the Standard Model indicates that the ratio of muon neutrinos plus antineutrinos to electron neutrinos plus antineutrinos should be 2,

$$R \equiv \frac{\nu_\mu + \bar{\nu}_\mu}{\nu_{\rm e} + \bar{\nu}_{\rm e}} = 2 \qquad \text{(Standard Model).}$$

- Instead, Super Kamiokande confirmed that the ratio R is only 64% of what is expected.

This result could be explained by *neutrino flavor oscillations:* if the muon neutrinos oscillate into another flavor, the observed ratio would be reduced below the expected value.

Detailed analysis suggests that the oscillation partner of the muon neutrino is *not the electron neutrino.*

- Hence the super-K results are *not directly applicable to the solar neutrino problem*.
- ν_μ is oscillating either with the tau neutrino or some other flavor of neutrino that does not undergo normal weak interactions but does participate in neutrino oscillations ("*sterile neutrinos*").
- The best fit to the data suggests a mixing angle close to maximal (a *large mixing angle solution*) and

$$\Delta m^2 \simeq 5 \times 10^{-4} - 6 \times 10^{-3} \, \mathrm{eV}^2.$$

- The *large mixing angle* indicates that the mass eigenstates are approximately *equal mixtures of the two weak flavor eigenstates*.

12.9.2 SNO Observation of Neutral Current Interactions

The Super Kamiokande results cited above indicate conclusively the *existence of neutrino oscillations* and thus of *physics beyond the Standard Model*.

- However, the oscillations *do not appear to involve the electron neutrino*.
- Thus the Super Kamiokande results cannot be applied directly to the solar neutrino problem.

However, a water Cherenkov detector in Canada has yielded information about neutrino oscillations that *is* directly applicable to the solar neutrino problem.

The *Sudbury Neutrino Observatory (SNO)* could detect neutrinos in the usual way by the Cherenkov light emitted from

$$\nu + e^- \to \nu + e^- \qquad \text{(elastic scattering)},$$

but it differed from Super-K in containing *heavy water*.

- Heavy water is important because it contains *deuterium*.
- In regular water, to produce sufficient Cherenkov light the ν energy has to be greater than about 5–7 MeV.
- Because deuterium (d) contains a weakly-bound neutron, it can undergo a breakup reaction:
  - Any flavor neutrino can initiate the reaction

    $$\nu + d \to \nu + p + n \qquad \text{(Neutral current)}\ ,$$

  - but only electron neutrinos can initiate

    $$\nu_{\mathrm{e}} + d \to e^- + p + p \qquad \text{(Charged current)}.$$

- Both of these reactions have much larger cross sections than elastic neutrino–electron scattering, so SNO could gather events at relatively high rates.
- The energy threshold could be lowered to 2.2 MeV, the binding energy of the deuteron.
- Because the *neutral currents are flavor blind,* $\nu + d \to \nu + p + n$ gives SNO the ability to see the *total neutrino flux of all flavors* coming from the Sun.

Table 12.2: Comparison of SNO results and Standard Solar Model predictions for solar neutrino fluxes. All fluxes are in units of 10^6 cm^2 s^{-1}.

SSM ν_e Flux	SNO ν_e Flux	SNO ν_e/SSM	SNO all flavors	SNO All/ SSM
5.05 ± 0.91	1.76 ± 0.11	0.348	5.09 ± 0.62	1.01

Because of its energy threshold, SNO sees primarily ^{8}B solar neutrinos.

- The initial SNO results *confirmed results from the pioneering solar neutrino experiments:*
- a *strong suppression of electron neutrino flux* is observed relative to that expected in the Standard Solar Model.
- Specifically, SNO found that only about $\frac{1}{3}$ of the expected ν_e were being detected.

However, SNO went further.

- By analyzing the flavor-blind weak neutral current events, it was possible to show that

> The total flux of *all neutrinos* in the detector was almost exactly that expected from the Standard Solar Model.

- Table 12.2 summarizes.

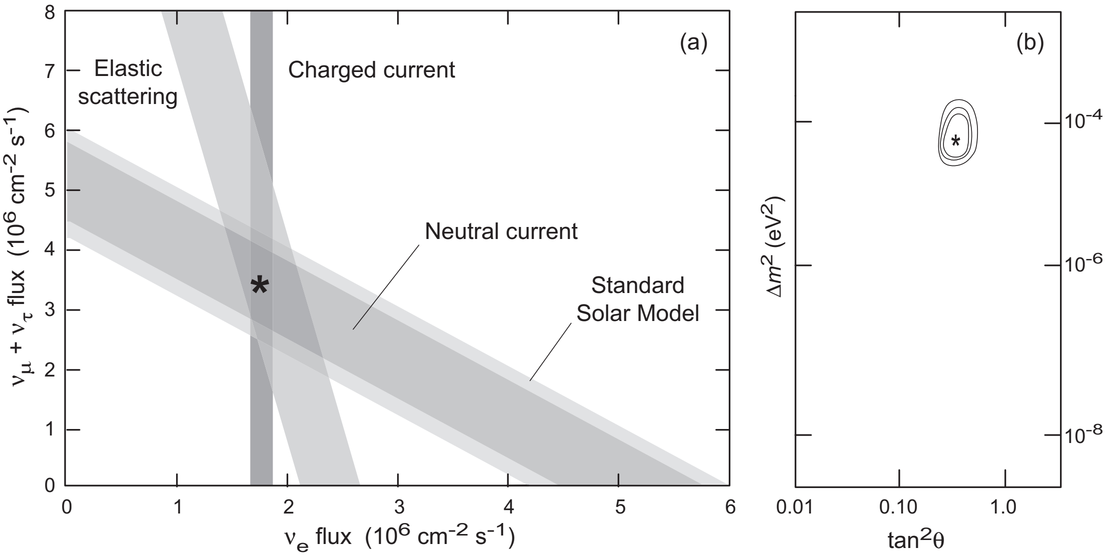
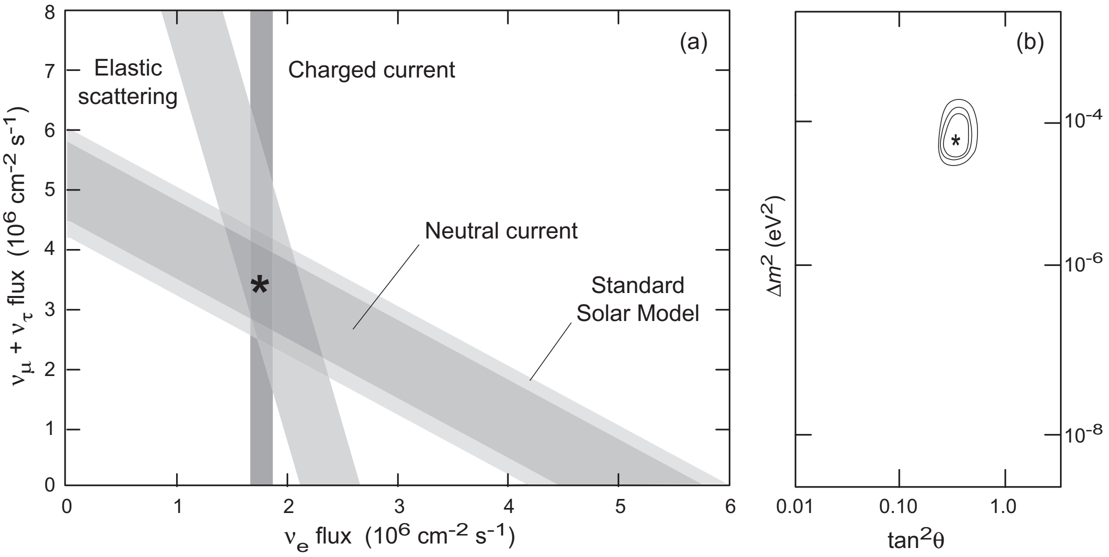


Figure 12.12: (a) Flux of solar neutrinos from ^{8}B detected for various flavors by SNO. The band widths represent one standard deviation. Bands intersect at the point indicated by the star, implying that about $\frac{2}{3}$ of the Sun's ^{8}B neutrinos have changed flavor between being produced in the Sun and being detected on Earth. The Standard Solar Model band is the prediction for the ^{8}B flux, irrespective of flavor changes. It tracks the neutral current band, which represents detection of all flavors of neutrino coming from the Sun. (b) 2-flavor neutrino oscillation parameters. The 99%, 95% and 90% confidence-level contours are shown, with the star indicating the most likely values. The best fit corresponds to the large-angle solution.

The SNO case for neutrino oscillation was strengthened by analysis of neutrino–electron elastic scattering data combined with SNO data from

$$\nu_e + d \rightarrow e^- + p + p \qquad \text{(Charged current).}$$

- Figure 12.12 illustrates.
- Best fit indicates that $\frac{2}{3}$ *of the Sun's electron neutrinos have changed flavor when they reach the Earth.*



Assuming a two-flavor mixing model, it is common to plot confidence level contours in a two dimensional plane with Δm^2 on one axis and $\tan^2\theta$ on the other.

- The figure above right shows the best-fit confidence-level contours for parameters based on SNO data.
- The SNO results suggest that the solar neutrino problem is solved by neutrino oscillations between ν_e and ν_μ flavors, with parameters

$$\Delta m^2 = 6.5^{+4.4}_{-2.3} \times 10^{-5}\,\mathrm{eV}^2 \qquad \theta = 33.9^{+2.4^\circ}_{-2.2^\circ}.$$

- This is again a *large-mixing-angle solution*, implying that
  - a ν_e is actually almost a strong superposition of two mass eigenstates,
  - probably separated by no more than a few hundredths of an eV.

12.9.3 KamLAND Constraints on Mixing Angles

KamLAND is housed in the same Japanese mine cavern that housed Kamiokande, predecessor to Super-K.

- It uses phototubes to monitor a large container of liquid scintillator.
- It looks specifically for electron antineutrinos produced during nuclear power generation in a set of 22 Japanese and Korean reactors that are located within a few hundred kilometers of the detector.
- The antineutrinos are detected from the inverse β-decay in the scintillator:

$$\bar{\nu}_{\rm e} + p \rightarrow e^+ + n.$$

- From power levels in the reactors, the expected antineutrino flux at KamLAND could be modeled.
- The experiment detected a shortfall of antineutrinos, consistent with a large-angle neutrino oscillation solution,

$$\Delta m^2 = 7.58^{+0.14}_{-0.13} \times 10^{-5}\,\text{eV}^2 \qquad \tan^2\theta = 0.56^{+0.10}_{-0.07},$$

corresponding to $\theta \sim 36.8^\circ$ for the vacuum mixing angle.

Combining the solar neutrino and KamLAND results leads to a solution

$$\Delta m^2 = 7.59 \pm 0.21 \times 10^{-5}\,\mathrm{eV}^2 \qquad \tan^2\theta = 0.47^{+0.06}_{-0.05},$$

- This implyies a vacuum mixing angle $\theta \sim 34.4^\circ$,
- which is again a *large mixing angle solution*.
- (Recall that θ has been defined so that its largest possible value is 45°.)

SNO and KamLAND results together greatly shrink the parameter space for solar neutrino oscillation parameters.

- Large mixing angle solutions were found by SNO and KamLAND for the ν_{e}–ν_μ mixing.
- This indicates that the vacuum oscillations of solar neutrinos are of secondary importance to the MSW matter oscillations in the body of the Sun itself.
- The inferred vacuum oscillation lengths for the large-angle solutions are much less than the Earth–Sun distance.
- Thus they would largely wash out any energy dependence of the neutrino shortfall.
- Since the detectors indicate that such an energy dependence exists, *the MSW resonance is strongly implicated as the source of the neutrino flavor conversion* responsible for the “solar neutrino problem”.

Ironically, the MSW effect

- was proposed to justify a *small mixing angle solution* but instead
- resolves the solar neutrino anomaly through a *large mixing angle solution*.

Chapter 13

Evolution of Lower-Mass Stars

Life on the main sequence is characterized by the stable burning of hydrogen to helium under conditions of hydrostatic equilibrium.

- While the star is on the main sequence the inner composition is changing, but there is little outward evidence until about 10% of the hydrogen is exhausted.
- Then the star experiences a (relatively) rapid series of changes that take it away from the main sequence.
- Stellar evolution after the main sequence is of short duration relative to the main sequence.
- However, post main-sequence evolution is generally more complex than main-sequence evolution.

Accordingly, let us now turn to a discussion of post main-sequence evolution.

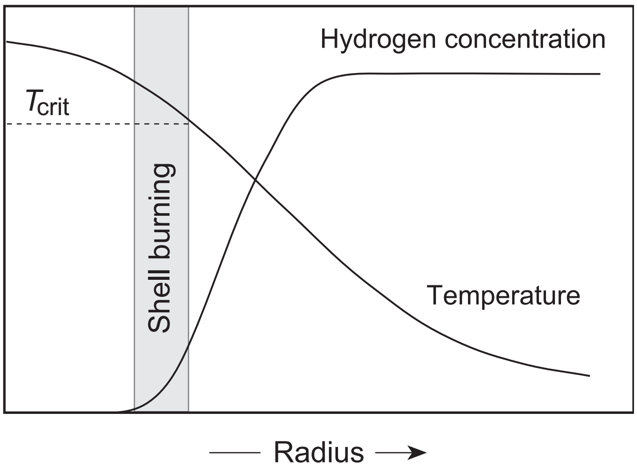


Figure 13.1: Conditions for hydrogen shell burning.

13.1 Shell Burning

An important aspect of post main-sequence evolution is the establishment of *shell burning sources* (Fig. 13.1).

- As the initial core hydrogen is depleted, a thermonuclear ash of helium builds up in its place.
- This ash is inert at hydrogen fusion temperatures because much higher temperatures and densities are necessary to initiate helium fusion.
- However, as the core becomes depleted in hydrogen there remains a concentric shell in which the hydrogen concentration and the temperature are both sufficiently high to support hydrogen fusion (Fig. 13.1).

This is termed a *hydrogen shell source.*

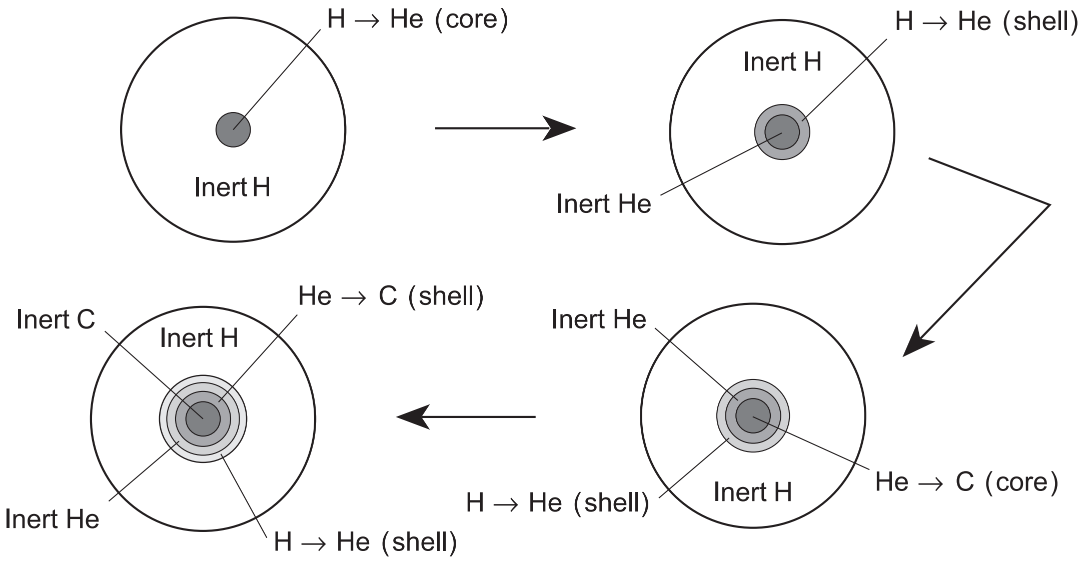


Figure 13.2: Schematic illustration of successive shell burnings.

As the core contracts after exhausting its hydrogen, the temperature and density rise and ignite helium in the core.

- As helium burns in the core a central ash of carbon is left behind that is inert because much higher temperatures are needed to fuse it to heavier elements.
- This is termed core helium burning.
- Just as for hydrogen, once sufficient carbon ash has accumulated in the core, helium burning will be confined to a concentric shell around the inert core.
- this is termed a *helium shell source.*

If the star is massive, this scenario may be repeated for heavier core and shell sources (Fig. 13.2).

The shell and core sources described above are not necessarily mutually exclusive.

- For more massive stars there may exist at any particular time

  - only a core source,
  - only a shell source,
  - multiple shell sources, or
  - a core source and one or more shell sources

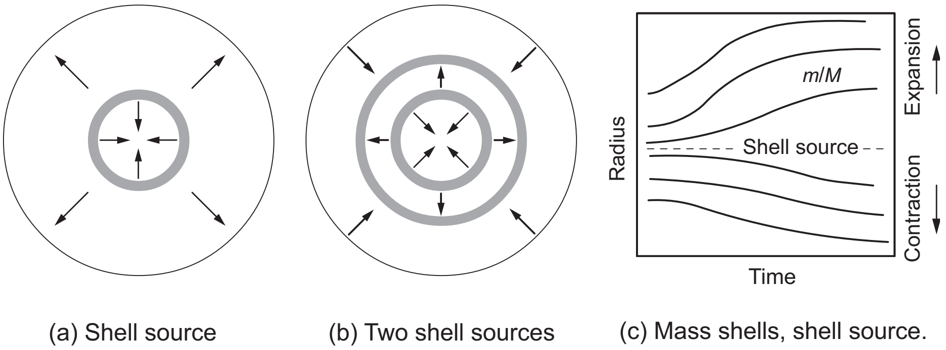
  burning simultaneously.

- These sources can have complicated instabilities and interactions.

> An important concept for understanding the action of shell sources is the *mirror principle*, which we now describe.

Mirror Response of Mass Shells: An important aspect of shell energy sources is termed the *mirror principle*.

- Experience with simulations indicates that *shell sources tend to produce "mirror" motion of mass shells above and below them,* as illustrated in the figure below.



(a) Shell source (b) Two shell sources (c) Mass shells, shell source.

- If there is a single shell source the mass layers below the shell source tend to contract and the mass layers above the shell source tend to expand, as illustrated in (a) and (c).
- For two shell sources, each tends to mirror the mass shells above and below, as illustrated in (b).
- In the absence of core burning, with two shell sources the core tends to contract, so by the mirror principle the layers above the inner shell source tend to expand (moving the second shell source further outward).
- Applying the mirror principle to the outer shell source, the layers outside the outer shell source (surface layers, for example) will tend to contract.

Motion in the HR diagram in late stellar evolution simulations often can be predicted by this principle of mirrored motion.

13.2 Stages of Red Giant Evolution

Globular clusters have HR diagrams differing substantially from those for stars near the Sun or for open clusters.

- We have interpreted this provisionally as evidence that *globular clusters are old* and that
- these differences are connected with the *time evolution of star populations.*
- We are now in a position to place those qualitative remarks on a much firmer footing.
- The most distinctive features of the HR diagrams for old clusters are

  1. The absence of main-sequence stars above a certain luminosity, and
  2. Loci of enhanced populations in the giant region termed the

     - *red giant branch* (RGB),
     - the *horizontal branch* (HB), and
     - the *asymptotic giant branch* (AGB),

     respectively.

These are illustrated schematically in Fig. 13.3 and for an actual cluster in Fig. 13.4 (see next page).

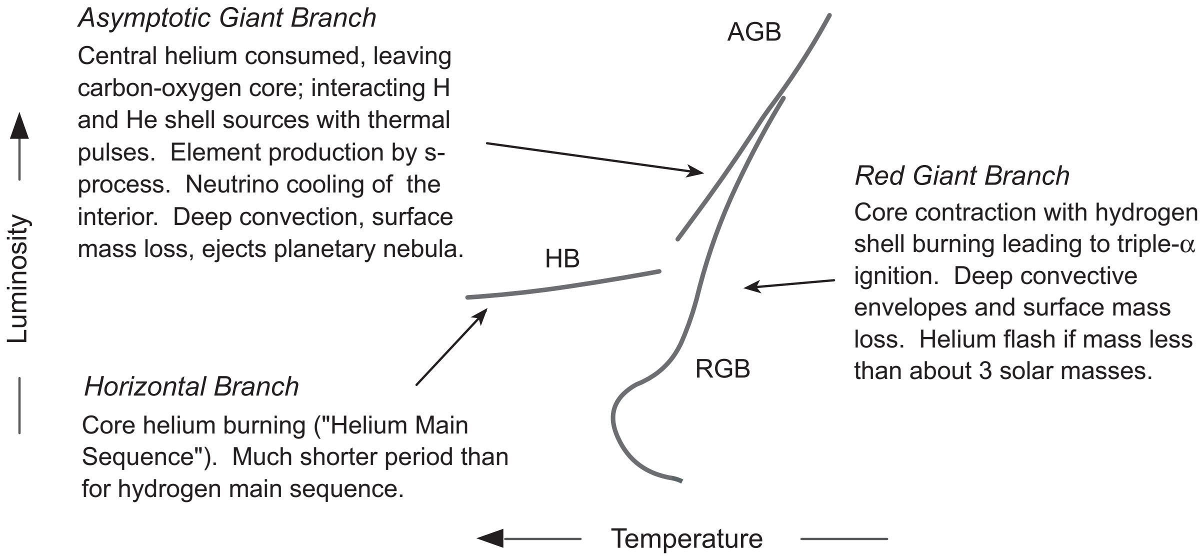

Asymptotic Giant Branch
Central helium consumed, leaving carbon-oxygen core; interacting H and He shell sources with thermal pulses. Element production by s-process. Neutrino cooling of the interior. Deep convection, surface mass loss, ejects planetary nebula.
AGB
Red Giant Branch
Core contraction with hydrogen shell burning leading to triple-α ignition. Deep convective envelopes and surface mass loss. Helium flash if mass less than about 3 solar masses.
HB
RGB
Horizontal Branch
Core helium burning ("Helium Main Sequence"). Much shorter period than for hydrogen main sequence.
Luminosity
Temperature

Figure 13.3: Schematic giant branches in an evolved cluster.

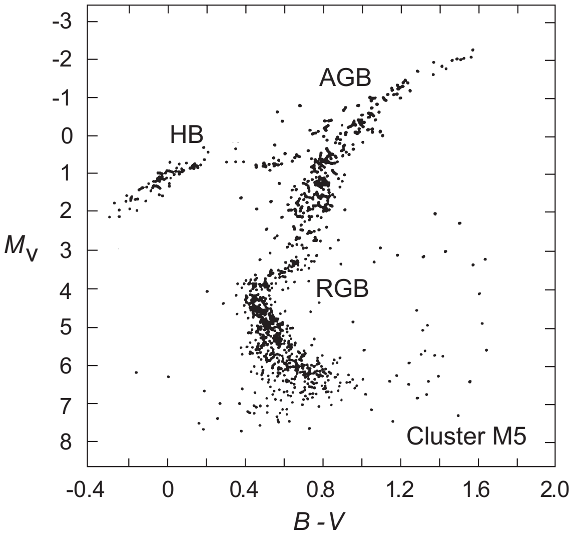


Figure 13.4: Actual giant branches for the globular cluster M5 in Serpens.

- Regions of enhanced population in the HR diagram are a signal that individual stars spend significant portions of their lives in these regions.
- As we now discuss, the
  - red giant branch,
  - horizontal branch, and
  - asymptotic branch

  can be identified with *distinct stages of post main-sequence evolution* for intermediate mass stars.
- These stages
  - are of short duration compared with main-sequence evolution, but
  - are *long compared with stages in between*.

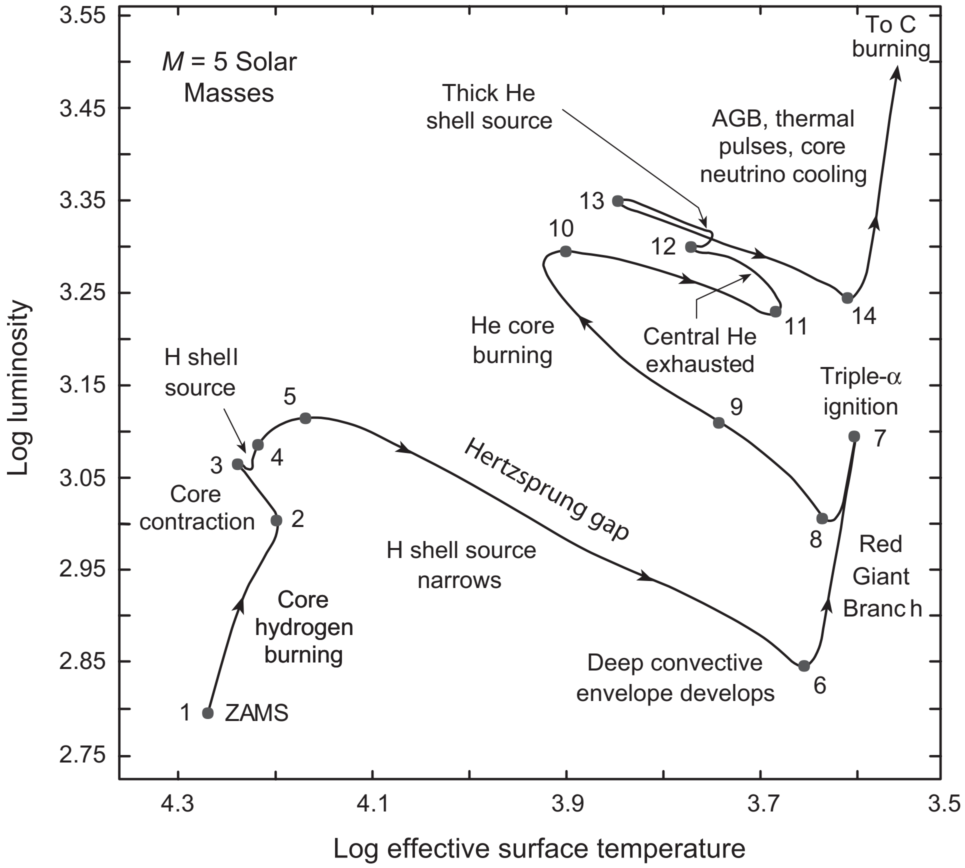


Figure 13.5: Evolution away from the main sequence for a 5 solar mass star.

As representative, we consider the calculated evolution of a 5 solar mass star, as illustrated in Figs. 13.5 and 13.6.

- Beginning at ZAMS the star converts hydrogen to helium.
- This causes a very small upward drift on the HR diagram.
- As core hydrogen is depleted the core contracts and eventually a hydrogen shell source is established.

These events signal the advent of a rapid departure from the main sequence that we will now follow in some detail.

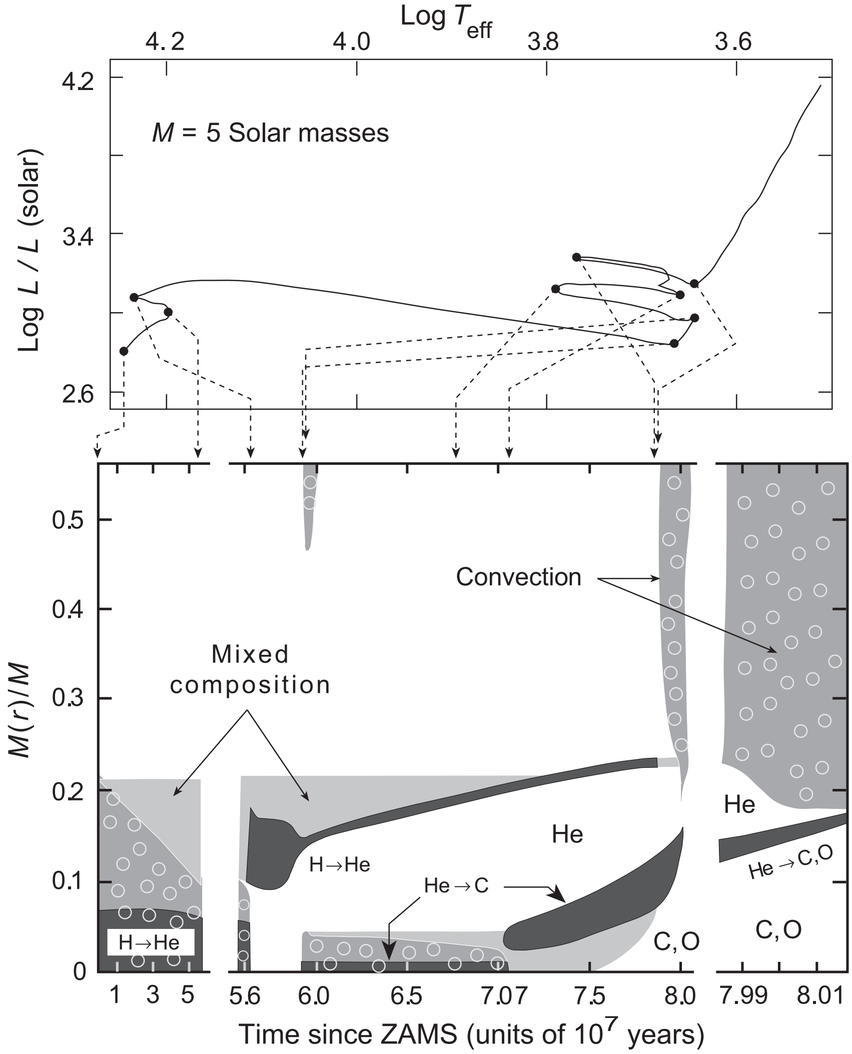


Figure 13.6: Post main-sequence evolution of a 5 solar mass star. Darkest shading indicates regions of energy production and regions with circles are convective. Note the breaks in scale for the time axis.

13.3 The Red Giant Branch

The hydrogen shell source established when the core hydrogen is depleted burns outward, leaving behind a helium-rich ash.

- The sole energy source is in a concentric shell, so
  - the core *cannot maintain a thermal gradient* and
  - it *equilibrates in temperature*.

> Such *isothermal cores* are characteristic of stars that have only shell energy sources.

- As the core increases in size because of the shell burning it is supported primarily by the pressure of the helium gas, which is typically still *nondegenerate* and *nonrelativistic*.
- But there is a limit to the mass of an isothermal helium core that can be supported by the gas pressure.
- This *Schönberg–Chandrasekhar limit* is given by

$$M_{\rm c} \simeq 0.37 \left(\frac{\mu_{\rm env}}{\mu_{\rm c}} \right)^2 M,$$

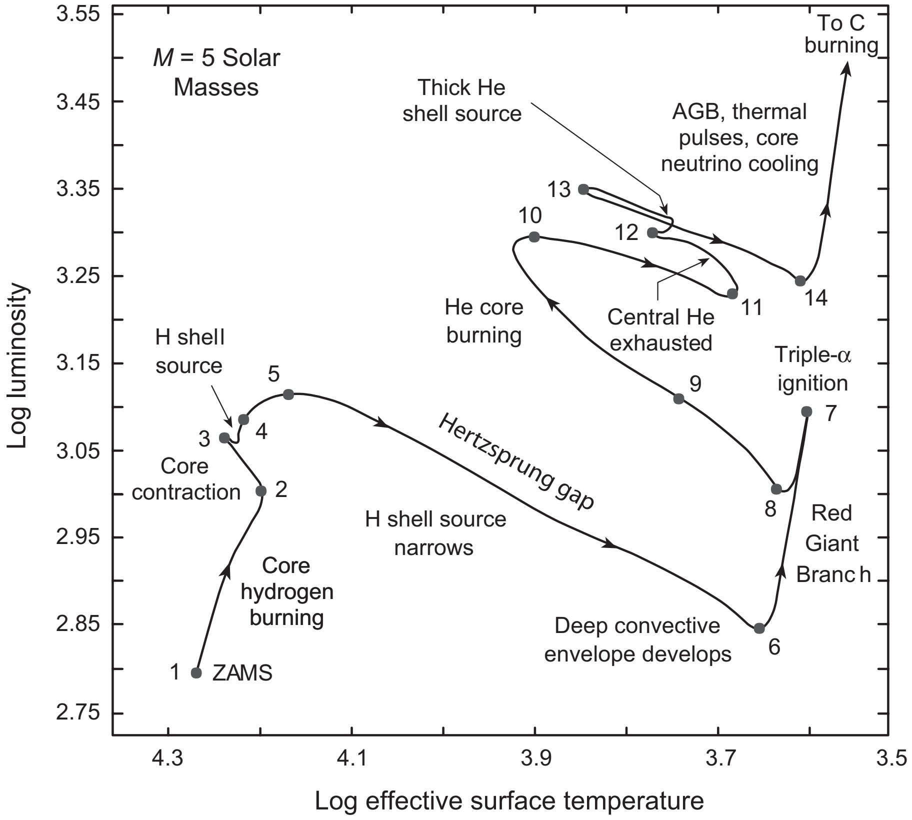
  for an isothermal core of ideal helium gas, where
  - M is the total mass of the star, $M_{\rm c}$ is the mass of the isothermal core,
  - $\mu_{\rm c}$ is the mean molecular weight in the core, and $\mu_{\rm env}$ is the mean molecular weight in the envelope.

Growth of an isothermal helium core to this size

- typically requires that about 10% of the original hydrogen be burned,
- which is one basis for the earlier statement that significant evolution from the main sequence commences when 10% of hydrogen has been consumed.

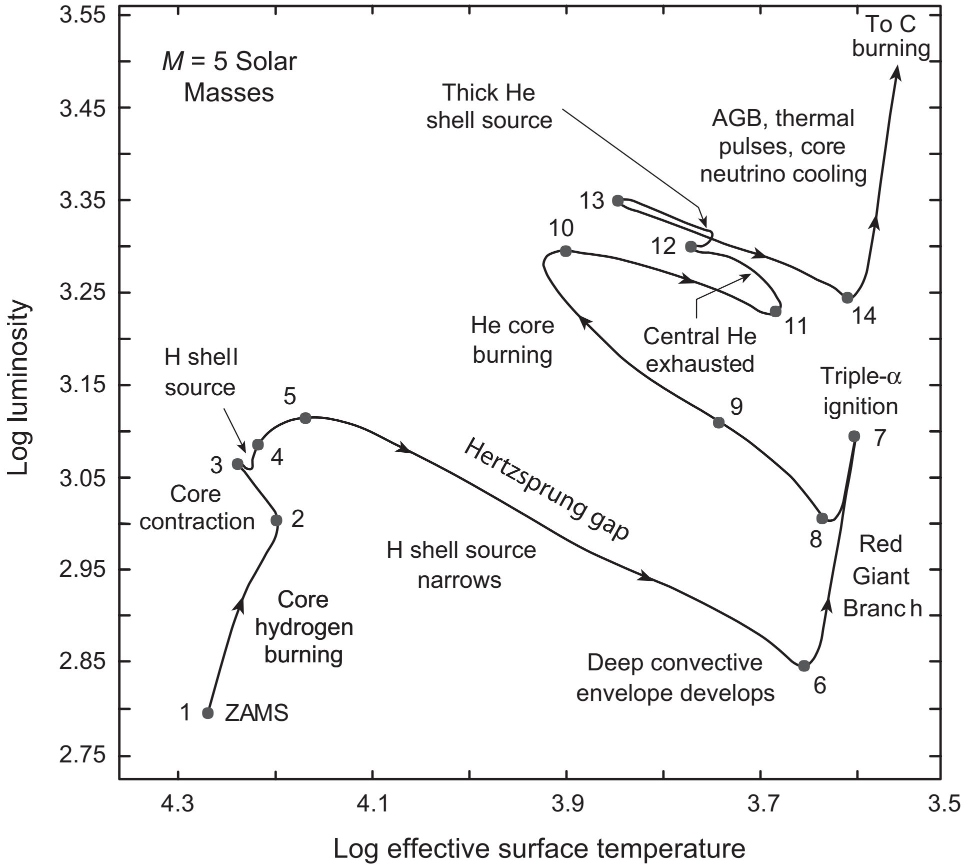
When the Schönberg–Chandrasekhar limit is reached the core can no longer support itself, or the layers above, against gravity.

- It begins to *contract* on a *Kelvin–Helmholtz timescale,*
  - which is *slow* compared to the dynamical timescale
  - but *rapid* compared to the nuclear burning timescale governing the time spent on the main sequence.
- The *contraction continues* until
  - *ignition of core helium burning* provides stabilizing pressure, or
  - until interior densities are reached where the *electron gas becomes degenerate*.
- Provided that the core mass does not exceed about $1.4\,M_\odot$,
  - degeneracy pressure stops the contraction,
  - but only after the core has become much hotter and denser, and
  - substantial gravitational energy has been released.
- Much of the energy released in the contraction is deposited in the envelope, which expands and cools, *enlarging* and *reddening* the photosphere.
- Thus the star evolves rapidly *upward and to the right* relative to the main sequence into the *red giant region* of the HR diagram.



- The region between the main sequence and the RGB (between points 5 and 6) has few stars (*Hertzsprung gap*).
- The star evolves so rapidly through this region that there is *little chance of observing it* (8×10^5 years; Table 13.1).
- As the envelope temperature decreases
  - *opacity increases* and
  - dT/dr becomes *steeper than the adiabatic gradient*.

Thus the star's envelope becomes *convective*.

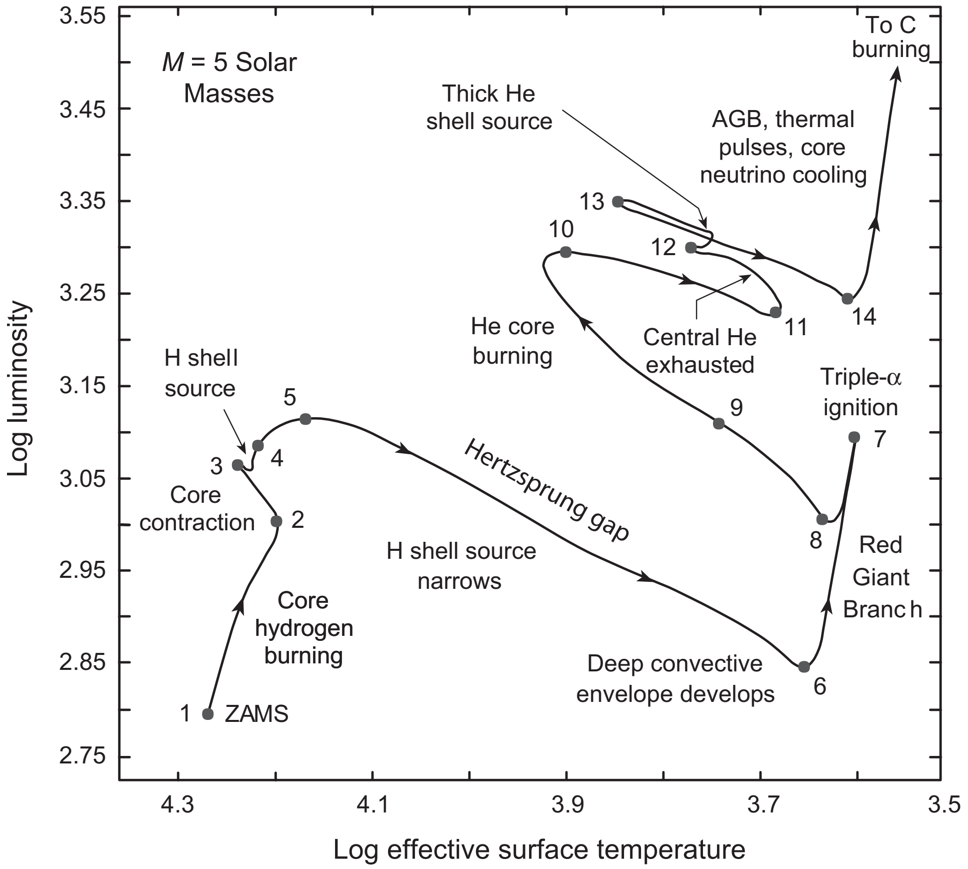


- We may then view the evolution to the red giant region as something like *the inverse of the collapse of fully convective protostars to the main sequence.*

- The almost fully convective star *climbs the Hayashi track in reverse* to the red giant region.

- The corresponding evolution in the above figure is on the red giant branch between the points labeled 6 and 7.

- While on the red giant branch the greatly-expanded star can exhibit *significant envelope mass loss*, with rates as large as $10^{-6} M_\odot$ per year observed for RGB stars.



13.4 Helium Ignition

- *Helium burning* by the *triple-α reaction* will be triggered when the core temperature reaches about 0.8×10^8 K.
- The *onset of helium burning* corresponds to the cusp shown in the above figure at point number 7, and signals the *end of RGB evolution*.
- Ignition of the core helium is *qualitatively different for stars above and below about* $3M_\odot$.

High mass stars generally have higher core temperatures than low mass stars at all stages of their evolution.

- Calculations indicate that stars of about $6M_\odot$ or more have high enough central temperatures to *evolve all the way to helium burning without their cores becoming degenerate*.
- Thus for $M >\sim 6M_\odot$, the initiation of core helium burning is likely a *smooth and orderly process*.
- But calculations indicate that for stars of about $3M_\odot$ or less the core electrons will have become *highly degenerate before the triple-α sequence ignites.*
- The *equations of state for ideal gases and degenerate gases differ fundamentally* in the relationship between temperature and pressure:
  - For an ideal gas the pressure is proportional to temperature.
  - For a degenerate gas the *pressure is essentially independent of the temperature*.

13.4.1 Thermonuclear Runaways under Degenerate Conditions

Ignition of thermonuclear reactions under degenerate conditions leads to violent energy releases:

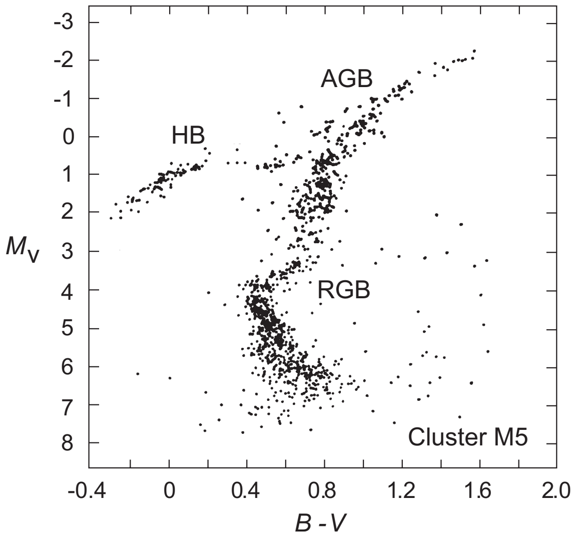
1. Ignition of the fusion reaction releases large amounts of energy, which quickly raises the local temperature.
2. In a normal explosion (ideal gas), a rise in temperature causes a corresponding rise in pressure that separates and cools the reactants, limiting the explosion.
3. Not so in degenerate gases because pressure is not increased initially by the sharp temperature rise.
4. Since charged-particle fusion reactions have very strong temperature dependence, the rise in temperature causes a rapid increase in the reaction rates and the fusion reactions run faster.
5. This in turn raises the temperature further and thus reaction rates, and so on *(thermonuclear runaway)*.
6. The large thermal conductivity of degenerate matter means a thermonuclear runaway triggered locally spreads rapidly through the degenerate matter.
7. This runaway continues until enough electrons are excited to lift the degeneracy of the electron gas.
8. The equation of state then tends to that of an ideal gas and the resulting increase of pressure with temperature moderates the reactions.

13.4.2 The Helium Flash

When such a thermonuclear runaway occurs under degenerate conditions for triple-α it is termed a *helium flash.*

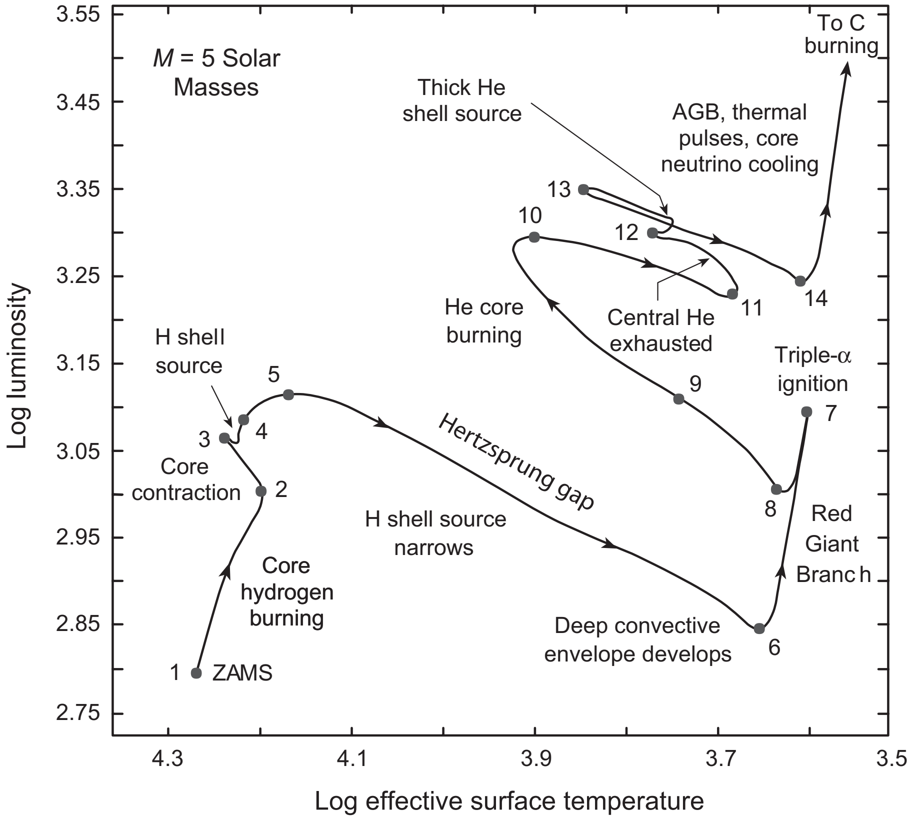
- Simulations show that stars of less than about $3M_\odot$ *ignite helium under degenerate conditions*.
- Simulations indicate further that the helium flash
  - ignites the entire core within seconds,
  - that the temperature can rise to more than 2×10^8 K before the runaway begins to moderate, and that
  - the energy release during the short flash can approach $10^{11} L_\odot$ *(comparable to the luminosity of a galaxy!).*
- However, this extremely violent event probably has little directly visible external effect because *the enormous energy release is almost entirely absorbed in the envelope.*
- In effect, the explosion is so strongly tamped by the external matter in the gravitational potential well of the star that it does not make it to the exterior.

- Once the degeneracy of the core is lifted following the helium flash (or following the onset of the triple-α reaction in heavier nondegenerate cores), the *helium burns steadily to carbon at a temperature of about* $1.5 \times 10^8\,K$.
- This signals the beginning of the *horizontal branch (HB)* portion of red giant evolution.



13.5 Horizontal Branch Evolution

- The horizontal branch (HB) of the above figure corresponds to a *period of stable core helium burning* that is in many ways analogous to the core hydrogen-burning main sequence.
- Thus, this period is often termed the *helium main sequence*.



- The HB corresponds to points 8–10 above.
- This *"helium-burning main sequence"* is a time of *hydrostatic equilibrium* for the same reasons as for the hydrogen-burning main sequence.
- The helium-burning main sequence is *much shorter than the hydrogen-burning main sequence,* in accord with earlier discussion of burning timescales.
- Initially on the HB the star typically has
  - lower luminosity than in the RGB period, but
  - higher than when it was on the main sequence.

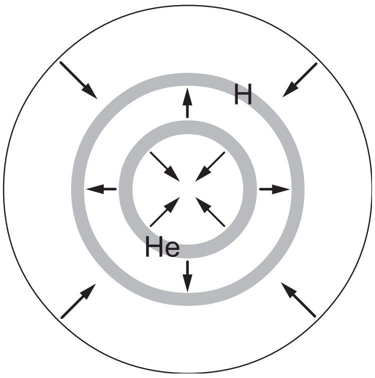
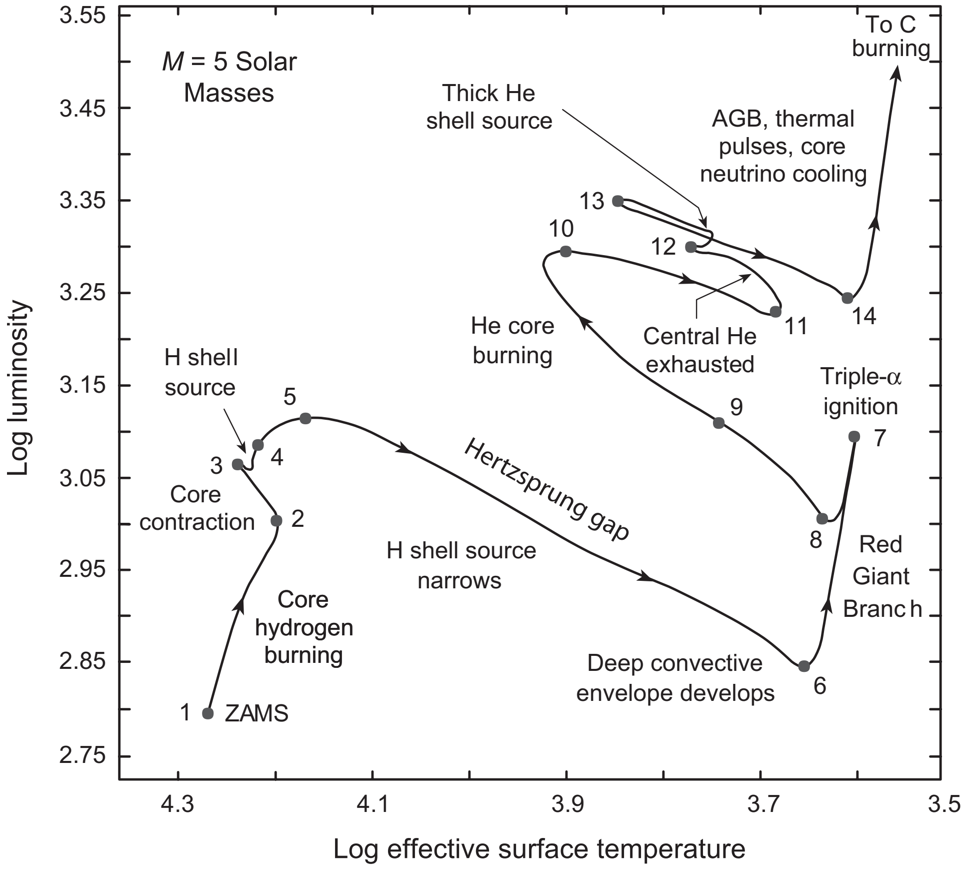
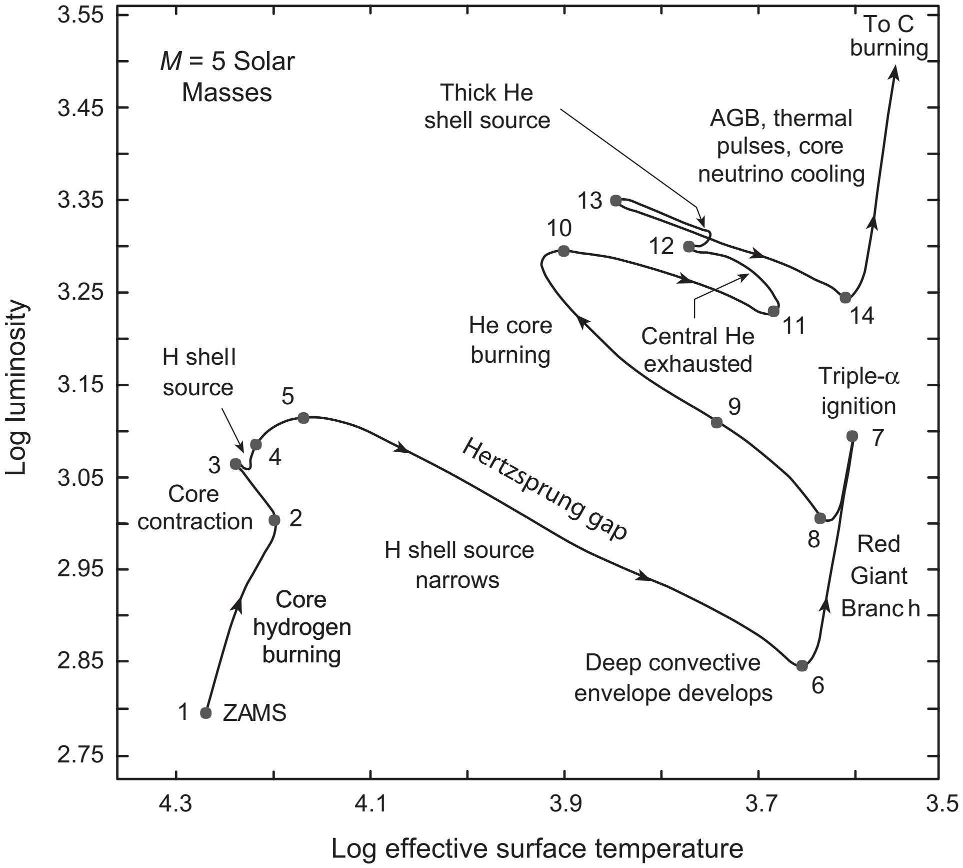


Figure 13.7: Mirror principle applied to helium and hydrogen shell sources in horizontal-branch evolution.

- The star remains on the HB while it has helium core fuel.
- When the core helium is exhausted, the core contracts and a thick helium burning shell is established.
- The star now has *two shell sources* (and no core source):
  - the *broad helium-burning shell* and
  - the *narrow hydrogen-burning shell* lying above it.
- Mirror principle (see Fig. 13.7):
  - the *inert carbon–oxygen core* inside the helium source *contracts* (no power source),
  - the *inert helium layer* outside the helium shell source *expands*, pushing the hydrogen shell source to larger radius, and
  - the *outer layers* of the star *contract*.



- The star *moves left on the HR diagram* and represents the evolution between points 11 and 13 in the above figure.
- The *helium shell source narrows and strengthens* as the core compresses further.
- Layers above the He shell source *expand and cool*, which
  - *turns off the hydrogen shell source* above the helium shell source,
  - leaving only a *single active (He) shell source*.



- In accordance with the mirror principle, the star *contracts inside the helium source and expands outside it*, and
- *drifts quickly to the right in the HR diagram* until it reaches the vicinity of the Hayashi track (point 14 above).
- This signals the transition to the *asymptotic giant branch (AGB)*.

13.6 Asymptotic Giant Branch Evolution

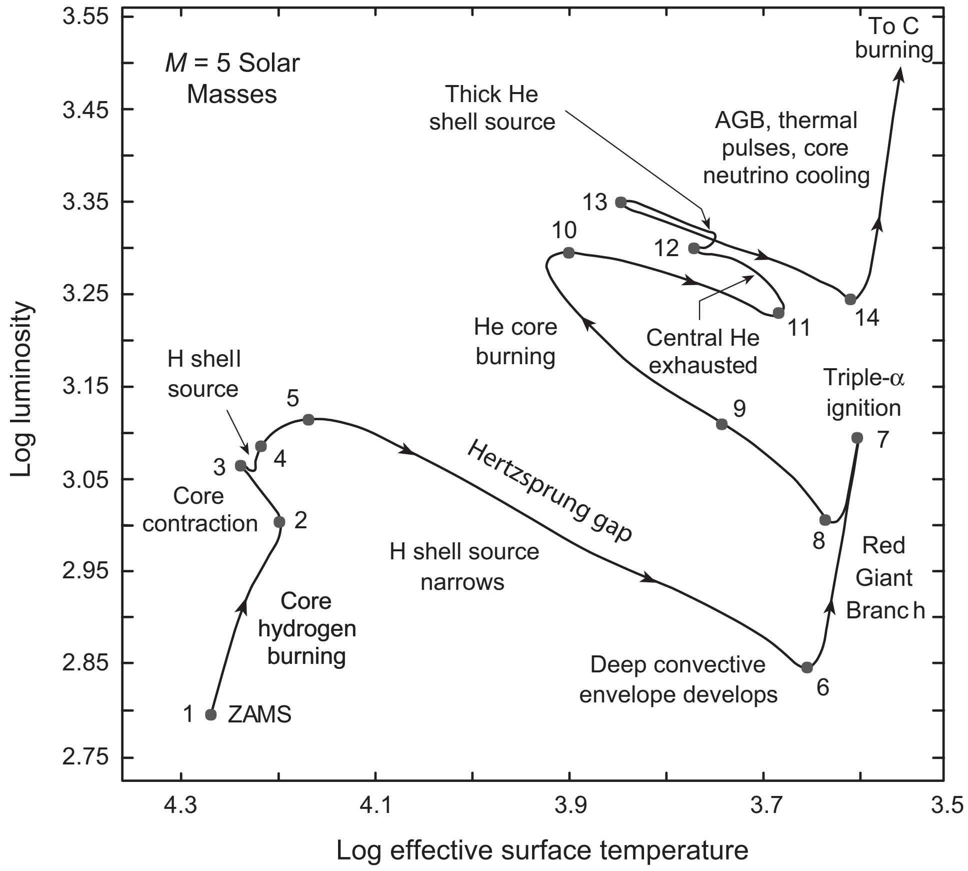
In many respects the evolution on the AGB now mimics that following the establishment of the first hydrogen shell source after core hydrogen was depleted on the main sequence.

- However, the star now has
  - an *electron-degenerate C–O core* and
  - *two shell sources* rather than one.

  (The hydrogen source turned off after ignition of the helium shell source but it will re-ignite after early evolution on the AGB.)

- The star again *increases in luminosity and radius* and
  - moves into the *red giant region* as earlier, but
  - at *even higher luminosities* on the AGB.

.



- The corresponding evolution in the above figure is from point 14 and beyond.
  - Roughly: continuation of the ascent on the RGB along the Hayashi track that was interrupted by ignition of the core helium.
  - If the star is massive enough, the *growing carbon core may ignite* eventually, but
  - if $M < 4-5M_{\odot}$ this is not likely and *all subsequent energy production will be from the shell sources*.

A number of important features characterize asymptotic giant branch evolution:

- The shell sources exhibit instabilities called *thermal pulses.*
- Shell sources in AGB stars are thought to be the primary site for the slow neutron capture or *s-process.*
- Stars in the giant region often exhibit *large surface mass loss*. Particularly true for AGB stars.
- *Deep convective envelopes* in the AGB phase can dredge elements synthesized in the interior up to the surface.
- These elements can then be *distributed to the interstellar medium by winds* from the surface.

Let's now discuss each of these important aspects of AGB evolution in more depth.

13.6.1 Thermal Pulses

The AGB period is characterized by the presence of both *hydrogen and helium shell sources.*

- However, these shell sources exhibit
  - instabilities
  - a complex interrelationship

  such that at any one time often only one of the two shell sources is burning.
- These instabilities are called *thermal pulses* or *helium shell flashes.*
- It will be shown below on rather general grounds that *a thin shell source is inherently unstable:*
  - Basically one finds that if shell sources are narrow enough the temperature *increases* upon expansion.
  - This is strongly destabilizing and sets the stage for a thermonuclear runaway.

Therefore, in many respects *a thin shell source behaves like a degenerate gas* with regard to thermal stability, even if the gas in the shell source is nondegenerate.

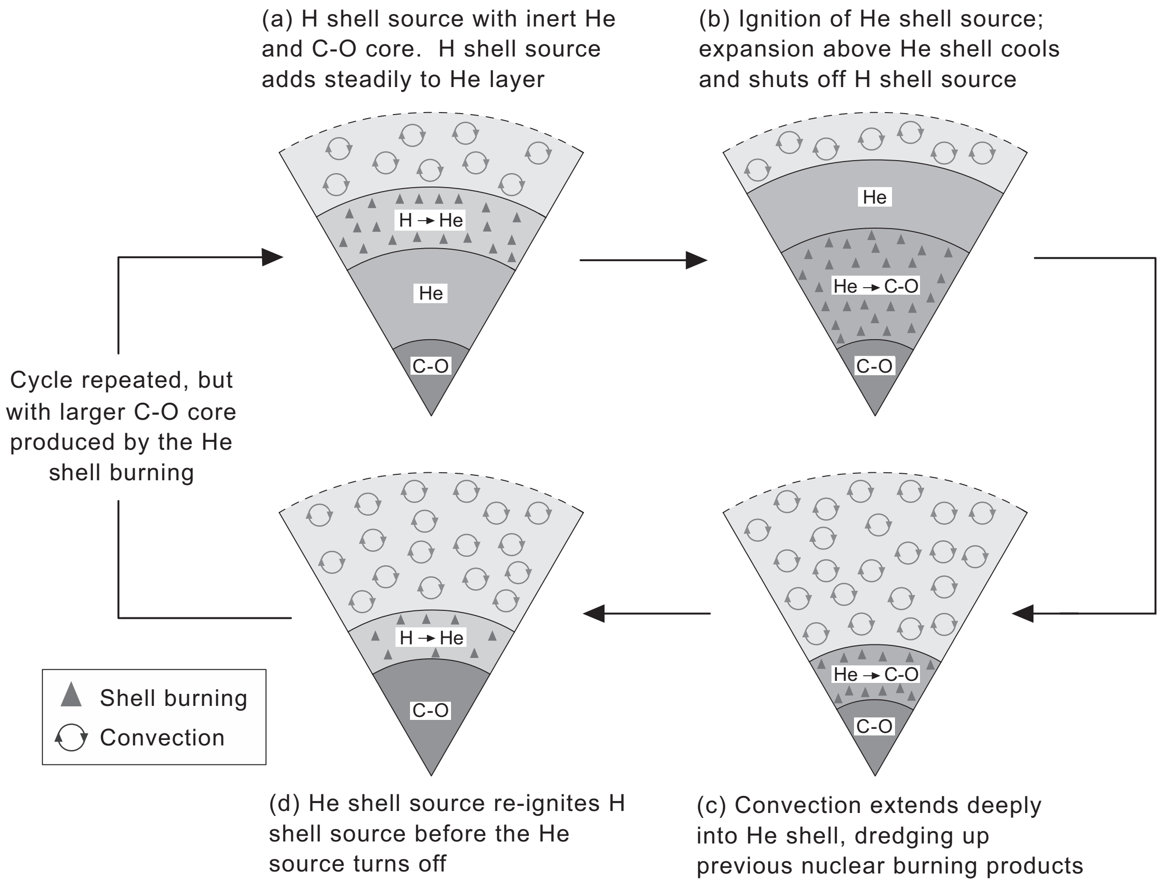
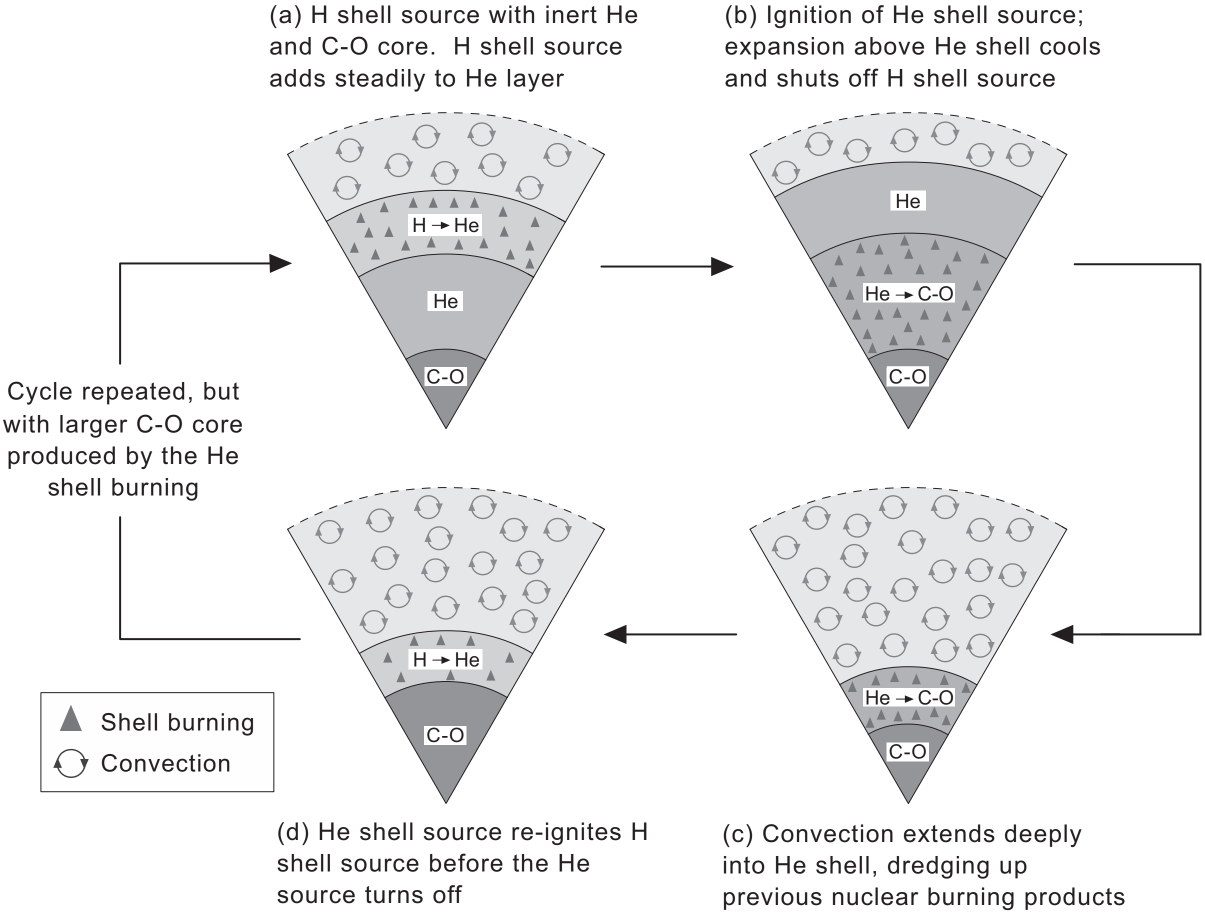

(a) H shell source with inert He and C-O core. H shell source adds steadily to He layer
H → He
He
C-O
(b) Ignition of He shell source; expansion above He shell cools and shuts off H shell source
He
He → C-O
C-O
Cycle repeated, but with larger C-O core produced by the He shell burning
▲ Shell burning
Convection
H → He
C-O
He → C-O
C-O
(d) He shell source re-ignites H shell source before the He source turns off
(c) Convection extends deeply into He shell, dredging up previous nuclear burning products

Figure 13.8: Schematic illustration of thermal pulses in an AGB star.

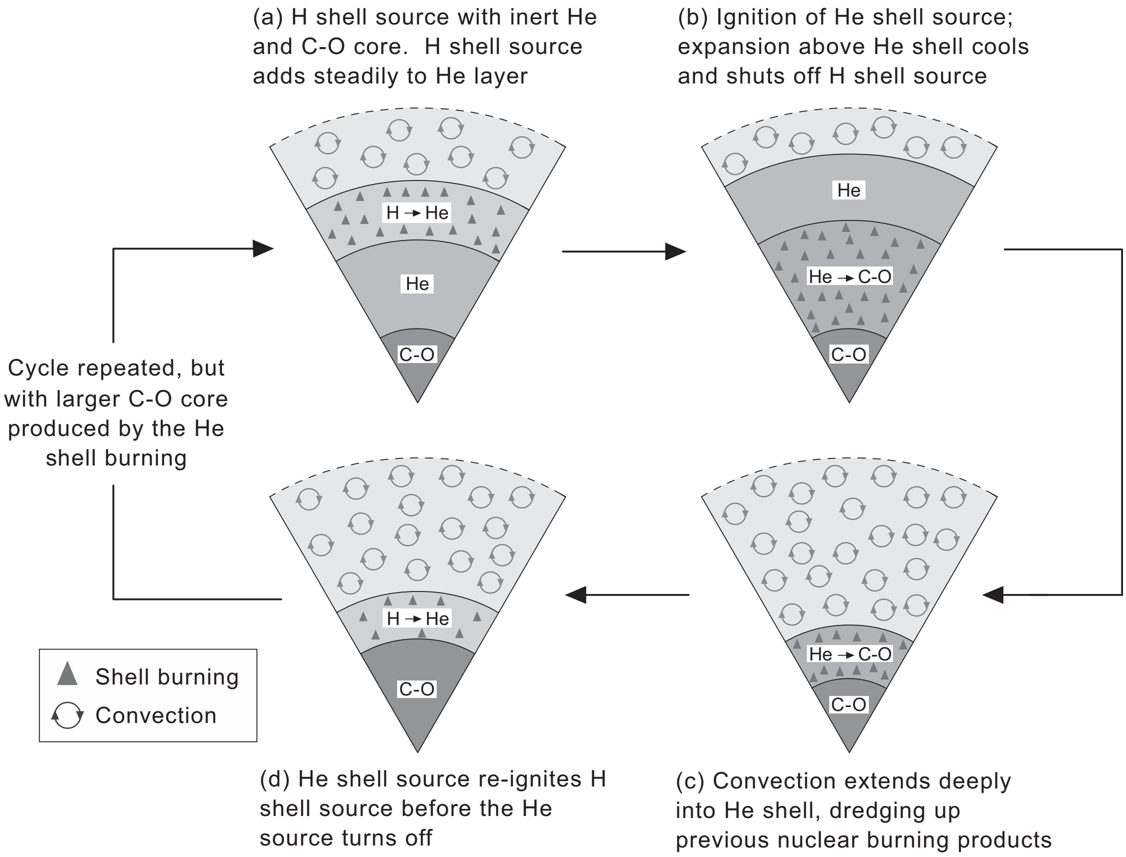
AGB *thermal pulses* are illustrated in Fig. 13.8.

- Let us assume that we have initially an
- inert C–O core surrounded by an *inert He layer*,
- with a *hydrogen shell source* at the base of the hydrogen layer above adding to the He layer (Fig. 13.8(a)).


(a) H shell source with inert He and C-O core. H shell source adds steadily to He layer
H → He
He
C-O
(b) Ignition of He shell source; expansion above He shell cools and shuts off H shell source
He
He → C-O
C-O
Cycle repeated, but with larger C-O core produced by the He shell burning
Shell burning
Convection
H → He
C-O
He → C-O
C-O
(d) He shell source re-ignites H shell source before the He source turns off
(c) Convection extends deeply into He shell, dredging up previous nuclear burning products

- As the core compresses the base of the helium layer may ignite, giving an *inner He shell source and an outer H shell source*.

- Expansion of layers above the hot He shell source lowers the temperature enough at the base of the hydrogen envelope to *turn off the H shell source*, leaving the star with a *single He shell source* (Fig. (b)).



(a) H shell source with inert He and C-O core. H shell source adds steadily to He layer

(b) Ignition of He shell source; expansion above He shell cools and shuts off H shell source

(c) Convection extends deeply into He shell, dredging up previous nuclear burning products

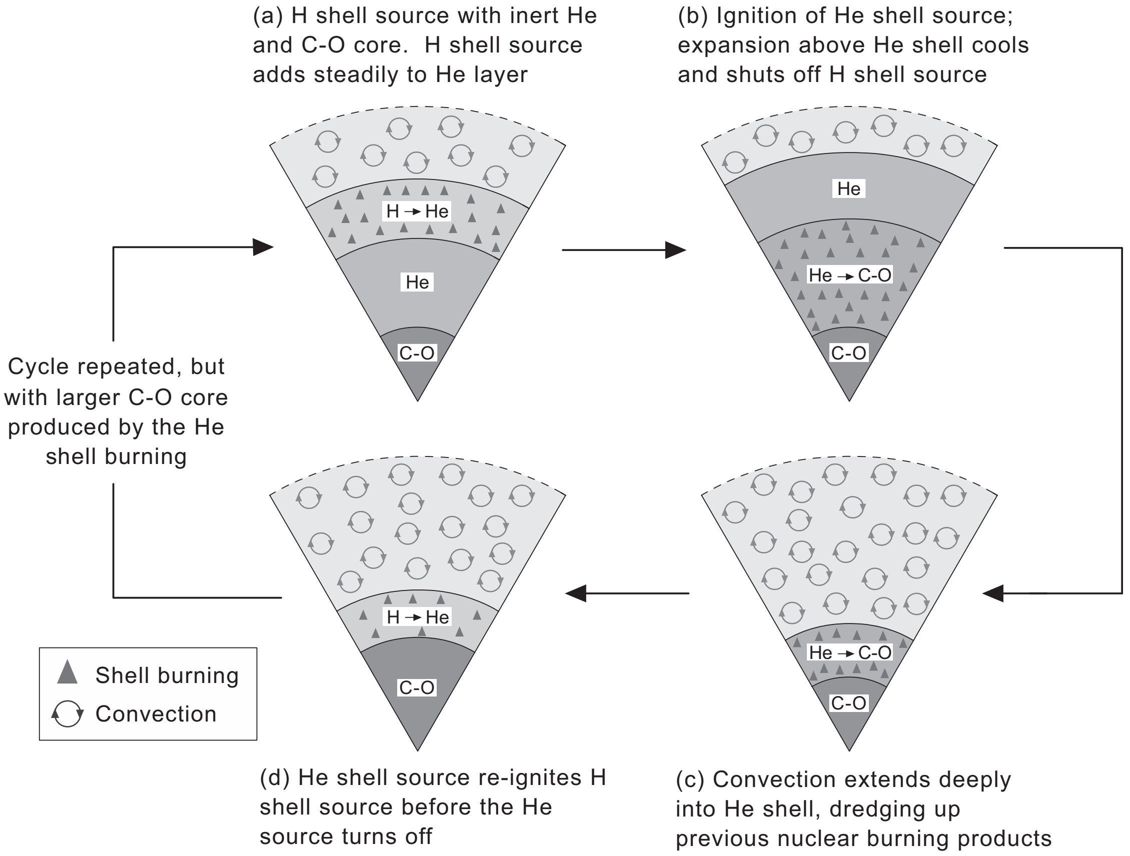
(d) He shell source re-ignites H shell source before the He source turns off

H → He
He
C-O
He
He → C-O
C-O
Cycle repeated, but with larger C-O core produced by the He shell burning
H → He
C-O
He → C-O
C-O
▲ Shell burning
⟳ Convection

- The hot helium shell source produces a *steep temperature gradient* and *convection develops that reaches down to the vicinity of the He shell source*, as illustrated in Fig. (c).

- This convection *mixes burning products from earlier evolution into the surface layers*.

- The He shell source burns outward, leaving a *growing C–O core* behind.



(a) H shell source with inert He and C-O core. H shell source adds steadily to He layer

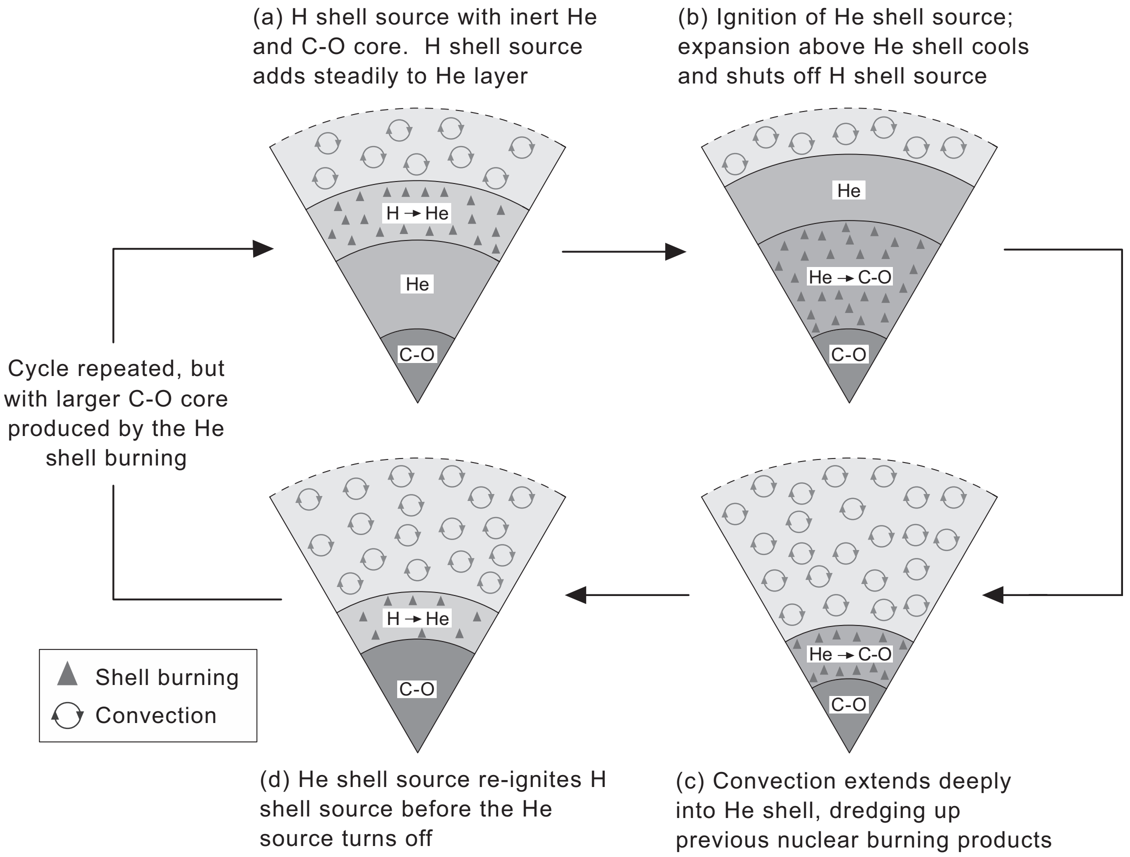
(b) Ignition of He shell source; expansion above He shell cools and shuts off H shell source

(c) Convection extends deeply into He shell, dredging up previous nuclear burning products

(d) He shell source re-ignites H shell source before the He source turns off

H → He
He
C-O
He
He → C-O
C-O
He → C-O
C-O
H → He
C-O
Cycle repeated, but with larger C-O core produced by the He shell burning
Shell burning
Convection

- The *He shell source eventually extinguishes* because of insufficient temperature at larger radius, but not before
- the *proximity of the hot He source re-ignites the shell source at the base of the hydrogen layer*, leading to the situation in Fig. (d).
- The *hydrogen shell source burns outward*, leaving behind a new layer of helium and *the cycle is repeated*,
- but now with a *larger carbon–oxygen core*.



(a) H shell source with inert He and C-O core. H shell source adds steadily to He layer

(b) Ignition of He shell source; expansion above He shell cools and shuts off H shell source

(c) Convection extends deeply into He shell, dredging up previous nuclear burning products

(d) He shell source re-ignites H shell source before the He source turns off

Cycle repeated, but with larger C-O core produced by the He shell burning

13.7 Stability of Thin Shell Sources

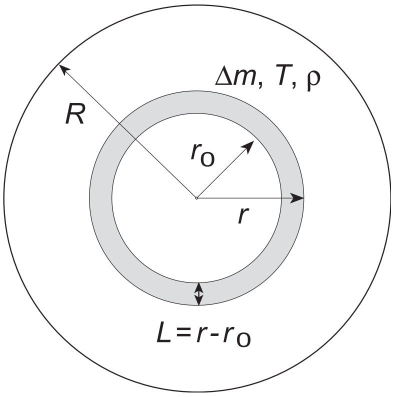
In the complicated tango between the shell sources that defines the cycle in the above figure,

- the *hydrogen shell source burns in stable fashion* but
- *the He shell source can be very unstable* because of
  - *strong temperature dependence* for He burning, and
  - because *thin-shell sources* are inherently *unstable*.

Consider a *thin shell source* in a star of radius R and mass M.

- The source is assumed to be in thermal equilibrium and to have
  - a mass Δm,
  - a density ρ,
  - a temperature T, and
  - a thickness $L = r - r_0 \ll R$,

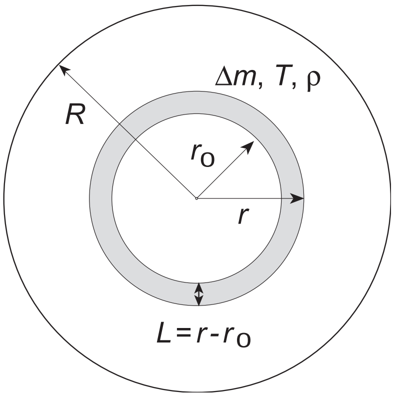
  as illustrated in the following figure.



If this is a single energy-producing shell in thermal equilibrium

- the shell is stable, with
- the *rate of energy flow* out of the shell = *rate of energy generation* in the shell.

We now consider the effect of a *fluctuation in energy* for this shell.



If the energy-generation rate is *increased by a fluctuation*

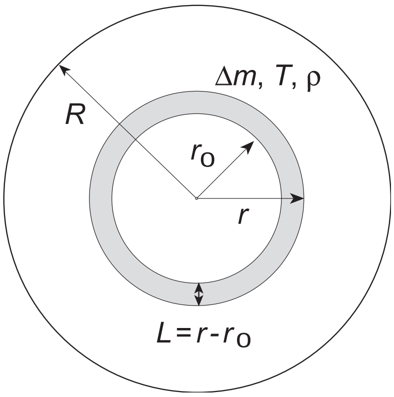
- the *shell will expand*, pushing the layers above it outward.
- Using the *generic equation of state*

$$\frac{dP}{P} = \alpha \frac{d\rho}{\rho} + \beta \frac{dT}{T}$$

with $\alpha \geq 0$ and $\beta \geq 0$,

- *stability of the shell* requires that $4L/r > \alpha$ (proof in Problem 13.4).
- But α is positive and finite, so for very thin shells $L/r \rightarrow 0$ and the *stability condition cannot be satisfied*.

This is called the *thin-shell instability;* it may be *significant for AGB stars* since they often develop thin shell sources.



Physically the thin-shell stability requirement

$$\frac{4L}{r} > \alpha$$

implies that

- if a shell source is narrow enough the temperature *increases* upon expansion,
- which is *strongly destabilizing* and sets the stage for a *thermonuclear runaway*.
- Therefore, in many respects *a thin shell source behaves like a degenerate gas* with regard to thermal stability, even if the gas in the shell source is not degenerate.

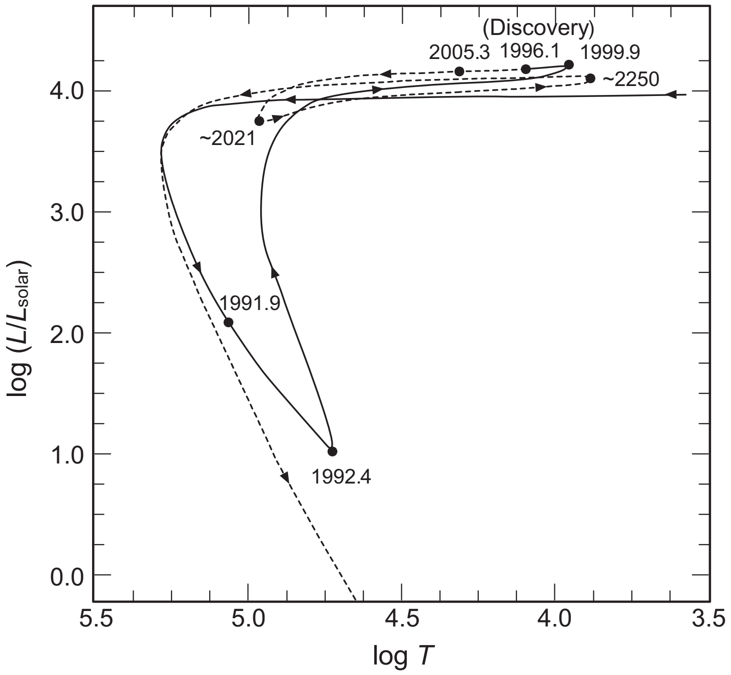
Thin He shell sources are particularly unstable because *helium is a very explosive thermonuclear fuel*.

Simulations indicate that an AGB star can undergo many thermal pulses (*tens to hundreds of pulses* before the envelope is eroded away by mass loss).

- The *thermal pulse durations* typically are only 10^4–10^5 years (a tiny fraction of the life of an average star).
- Thus it is very difficult to catch a star undergoing thermal pulses.
- About a quarter of AGB stars are predicted to undergo one final helium shell flash after hydrogen burning has ceased.
- This *late thermal pulse* occurs after the star has ejected most of its envelope as a planetary nebula and is settling into the white dwarf phase.
- Computer simulations of this event suggest that in such a star
  - the helium shell can re-ignite and the small remaining H envelope can be convectively mixed into the helium shell,
  - which leads to additional rapid hydrogen-driven flash burning and renewed mass ejection.
- Late thermal pulse events in asymptotic giant branch stars are expected to be rare, with a predicted rate of only about *one per decade in our galaxy.*

The star V4334 Sgr (*Sakurai's Object*) is thought to be a star caught undergoing a *late thermal pulse*.

- Since discovery in 1996, it has exhibited rapid evolution on the HR diagram accompanied by mass ejection.
- Model simulation of the evolution of Sakurai's object on the HR diagram is illustrated in the following



with a solid line indicating the prediction for evolution before discovery and a dashed line afterwards.

- These predicted loops imply surface-T variations by factors of 10 on timescales of 10–100 years.
- The observed surface T increased by about a factor of 2 in just the 10 years following discovery in 1996.

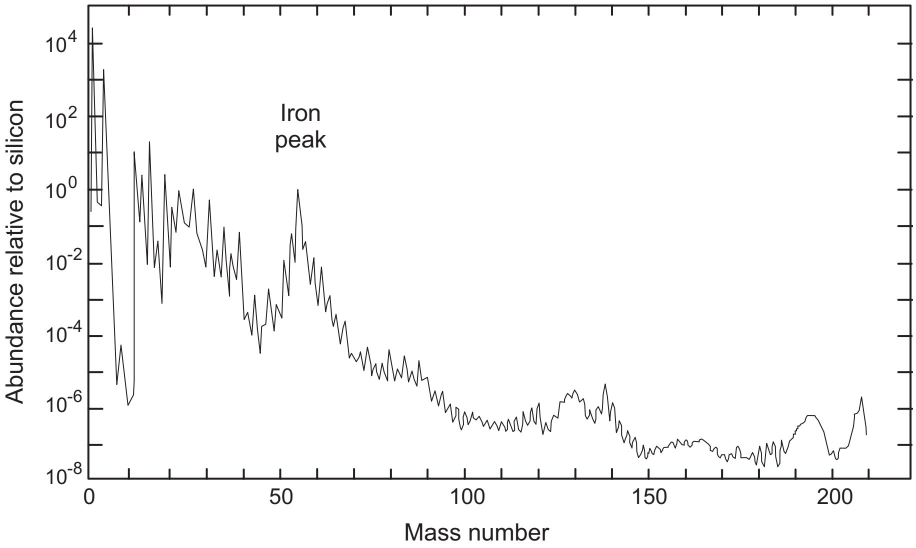


Figure 13.9: Solar System elemental abundances relative to silicon abundance.

13.7.1 Slow Neutron Capture

Figure 13.9 summarizes observed elemental abundances.

- *Elements up to iron* can be produced by fusion reactions and by nuclear statistical equilibrium in stars.
- *Elements beyond Fe* can't be produced in the same way because the Coulomb barriers become so large that extremely high temperatures are required.
- These high temperatures would produce a bath of high-energy photons that would *photodisintegrate any heavier nuclei that were formed*.
- Thus *other mechanisms produce heavier elements*.

One possibility is the *capture of neutrons on nuclei* to build heavier nuclei.

- Because neutrons are electrically neutral *they do not have a Coulomb barrier to overcome*.
- This permits reactions to take place at *low enough temperatures* that the newly-formed heavy nuclei will not be dissociated immediately by high-energy photons.
- There are *two basic neutron capture processes* that are thought to produce heavy elements:
  - the slow neutron capture or *s-process* and
  - the rapid neutron capture or *r-process*.
- Astrophysical sites for these neutron capture reactions have not been confirmed, but it is widely believed that
  - the *s-process* takes place in *AGB stars*
  - the *r-process* takes place in *neutron star mergers and in core-collapse supernova explosions*.

We shall discuss the s-process here and will address the r-process in later chapters.

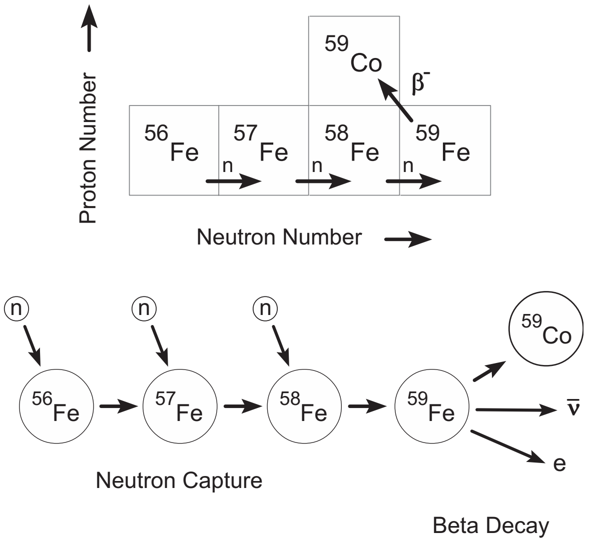
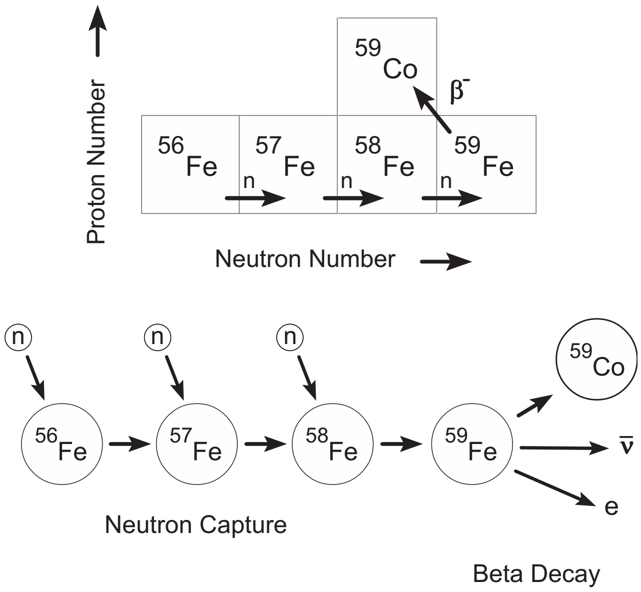


Figure 13.10: Example of slow neutron capture and β-decay (s-process).

The s-process refers to a *sequence of neutron capture reactions interspersed with beta decays* to produce heavier elements where the rate of *neutron capture is slow on a timescale set by competing beta decays*.

- We can illustrate by considering an iron nucleus subject to a low-intensity neutron source, as in Fig. 13.10.
- In this example, ^{56}Fe captures 3 neutrons sequentially to become ^{59}Fe.
- But as iron isotopes become neutron rich they become increasingly unstable against β^- decay.
- In this example the neutron flux is such that ^{59}Fe is likely to beta decay before it can capture another neutron.



- Now the ^{59}Co nucleus can *absorb neutrons and finally beta decay* to produce an isotope of the next atomic number (nickel), and so on.
- By this process, heavier elements can be *built up slowly* if a source of neutrons and the seed nuclei (iron in this case) are available.

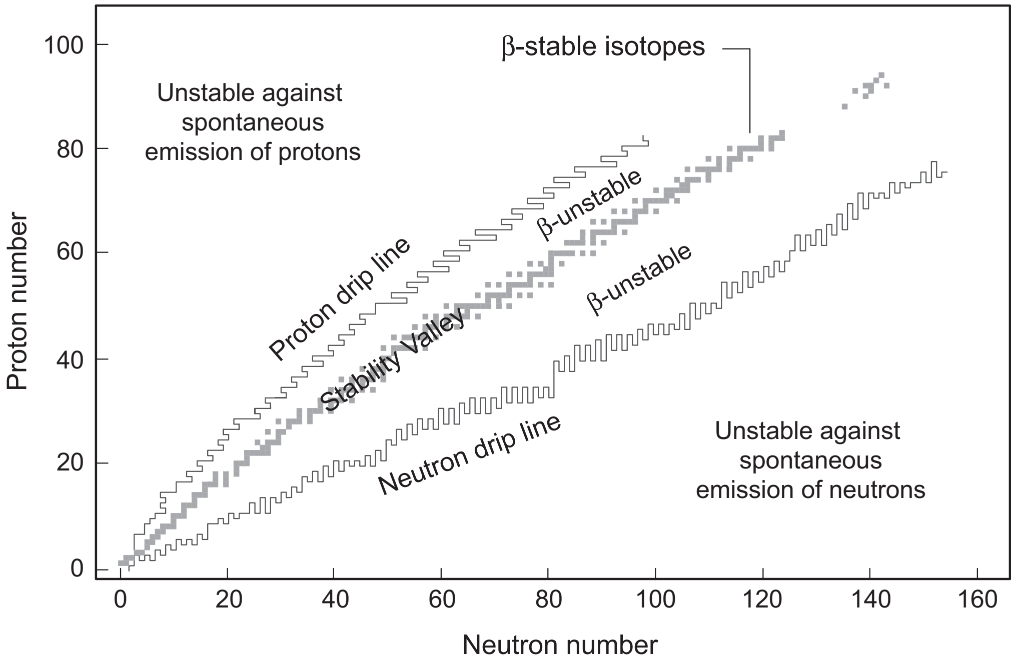


Figure 13.11: The valley of beta stability (shaded region). Isotopes lying in this valley are stable against β-decay. The "drip lines" mark the boundaries for spontaneous emission of protons or neutrons. Isotopes outside the stability valley are increasingly unstable against β-decay as one moves toward the drip lines.

- Because of the competition from beta decay, it is clear that the s-process can build new isotopes only in the *valley of beta stability* illustrated in Fig. 13.11.

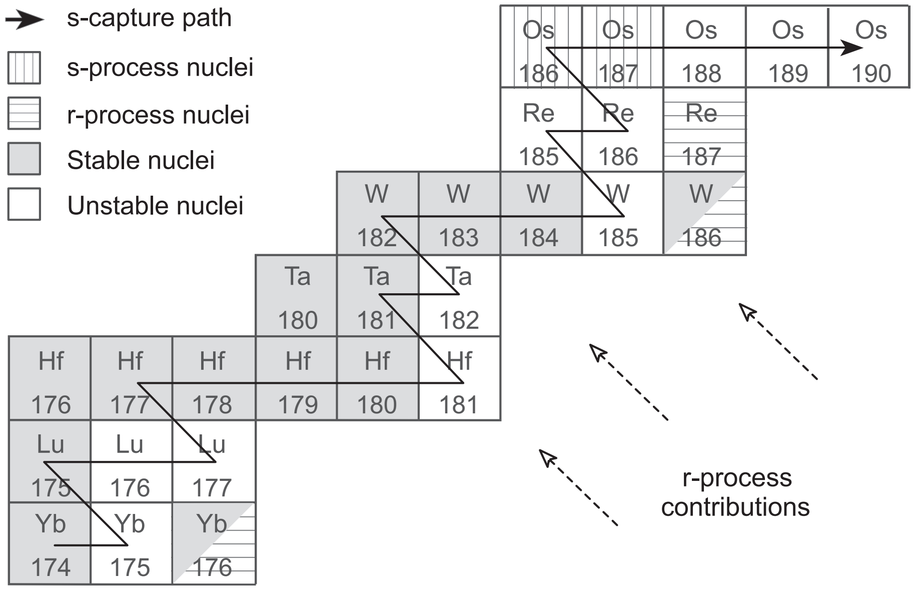


Figure 13.12: The s-process path in the Yb–Os region.

In Fig. 13.12 the *s-process path* is shown in the *Yb–Os region*.

- The path stays *very near the stability valley* (dark boxes).
- This figure also illustrates the *competition between the s-process and r-process* in producing the heavier elements.
- The r-process generally populates very neutron-rich isotopes that then β^- decay toward the stability valley.
- *Some isotopes (for example,* $^{186-187}$*Os) can be populated only by the s-process* because other stable isotopes protect them from r-process populations β-decaying from the neutron-rich side of the proton–neutron plane.
- *Other isotopes (e.g.* ^{186}*W) can be populated only by the r-process* because an unstable isotope lies to their left in Fig. 13.12, blocking the s-process path.

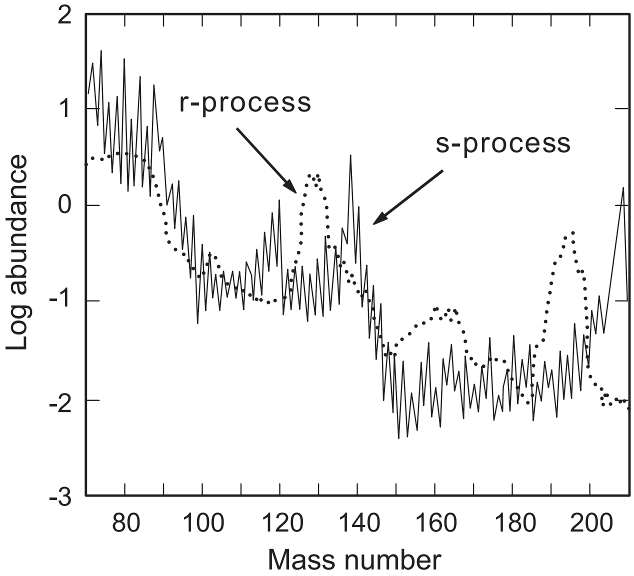


Figure 13.13: Relative contributions of the s-process and the r-process to heavy element abundances.

Many isotopes can be produced *by both the s-process and the r-process*. The *relative contributions of the s-process and r-process* to heavy element abundances are summarized in Fig. 13.13.

A source of slow neutrons is required for the s-process.

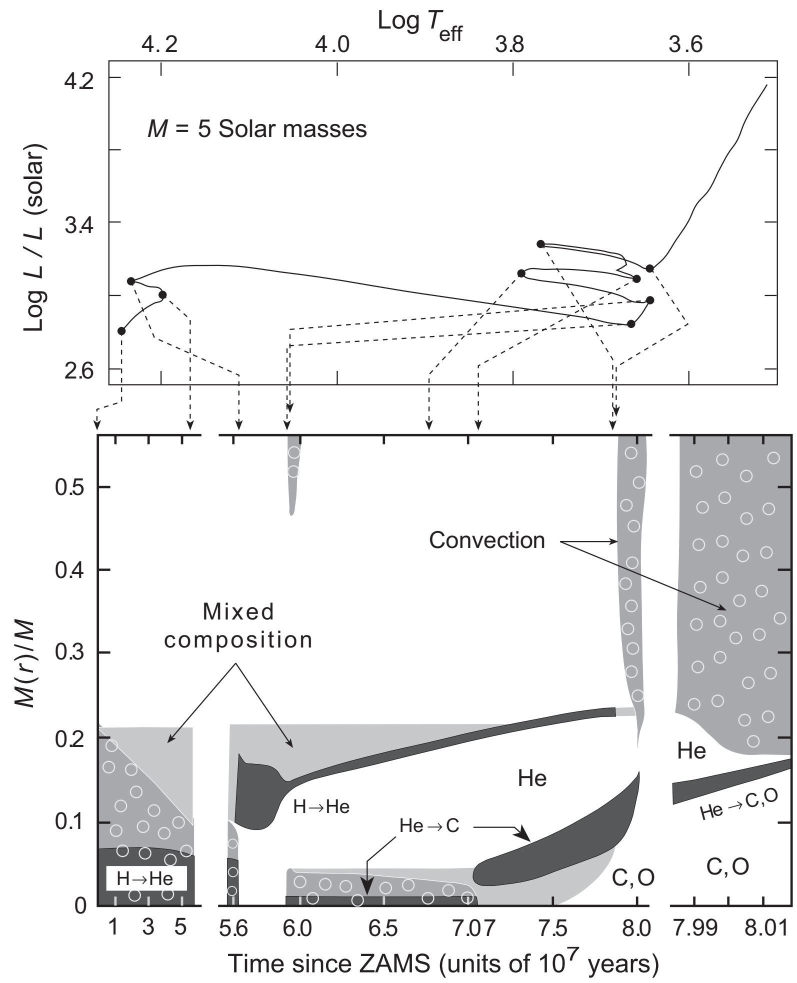
- Only a few nuclear reactions that are likely to occur in stars under normal conditions produce neutrons.
- *Free neutrons are unstable against β-decay* on a 10-minute timescale.
- Thus neutrons for the s-process are *not easy to come by*.

The box on the next page discusses possible neutron sources for the s-process that are thought to be present in red-giant stars during the AGB phase.

For the slow capture process it is thought that two reactions that can occur in AGB stars are primarily responsible for supplying the neutrons:

$$^{4}\mathrm{He} + {}^{13}\mathrm{C} \longrightarrow {}^{16}\mathrm{O} + \mathrm{n} \qquad {}^{4}\mathrm{He} + {}^{22}\mathrm{Ne} \longrightarrow {}^{25}\mathrm{Mg} + \mathrm{n}$$

- The $^{13}\mathrm{C}(\alpha,\,\mathrm{n})^{16}\mathrm{O}$ reaction is expected to provide the neutron flux at low neutron densities ($\lesssim 10^{7}\ \mathrm{cm}^{-3}$).
- The reaction $^{22}\mathrm{Ne}(\alpha,\,\mathrm{n})^{25}\mathrm{Mg}$ plays a secondary role, occurring at higher T during thermal pulses.



13.7.2 Development of Deep Convective Envelopes

- *Once a thin helium shell source develops* the resulting temperature gradients drive *very deep convection* extending down to the shell sources (above).
- As we shall discuss later, *mixing associated with this deep convection* is central to understanding the observation of nuclear-processed material associated with *surfaces and winds for red giant stars*.

13.7.3 Mass Loss

Observations indicate that once stars leave the main sequence they *experience large mass losses*, particularly in the *AGB and RGB phases*.

- This is most directly indicated by the observation of gas clouds with *outwardly directed radial velocities of 5–30 km s*$^{-1}$ near such stars.
- It has been found that this *mass loss is described by a semiempirical expression* of the form

  $$\dot{m} \simeq -A\frac{LR}{M}M_\odot\,\mathrm{yr}^{-1},$$

  where $A \sim 4\times10^{-13}$ is a constant, L is the luminosity, R is the radius, and M is the mass of the star.
- Thus, the *rate of mass ejection increases linearly* with
  - larger *luminosity*,
  - larger *radius*, and
  - smaller *mass*.
- This would be expected for mass loss from the surface of a luminous object with a surface gravitational field determined by its mass and radius.

Therefore, on the RGB and AGB

- the rapid increase in radius and luminosity leads to increased mass loss, and
- as the star sheds its matter the decreased residual mass reduces the gravitational potential and further accelerates the loss.
- The detailed mechanism is not well understood but:

> It is clear empirically that *mass loss can increase by orders of magnitude* relative to that associated with normal stellar winds in the *RGB and AGB phases*.

Example: Mass loss can be large for red giants.

- For RGB stars mass losses of $10^{-6}\,M_\odot\,\mathrm{yr}^{-1}$ have been recorded, while
- for AGB stars the losses can approach $10^{-4}\,M_\odot\,\mathrm{yr}^{-1}$.

If these rates were sustained a red-giant star would eject all of its mass on a timescale that is tiny compared to its overall lifetime.

13.8 Ejection of the Envelope

In the AGB phase the envelope of the star is consumed both from within and without:

- The surface is ejecting mass, while the carbon–oxygen core is growing internally as the shell sources burn outward.
- Detailed estimates indicate that the *surface mass loss is more important by orders of magnitude.*

This rapid loss of the envelope primarily from surface ejection while the core grows at very small comparative rates has two important implications:

1. The envelope of the star is lost rapidly into space, leaving behind a carbon–oxygen core.
   - The rapid loss of the envelope implies that a range of initial masses will leave behind cores (white dwarfs) of *almost the same mass.*
   - This is significant because white dwarf masses are observed to be concentrated in a narrow range near $M \simeq 0.6\, M_\odot$.
2. The ejected envelope is a natural candidate for producing *planetary nebulae,* which are commonly observed phenomena in late stellar evolution.

Thus, we expect the primary outcome of AGB evolution to be

- ejection of most of the star’s envelope as a *planetary nebula*,
- leaving behind a bare C–O (or Ne-Mg in heavier stars) core that will cool to form a *white dwarf*.

13.9 White Dwarfs and Planetary Nebulae

Late in the AGB phase, mass loss increases dramatically for a short period called the *superwind phase* (which, as for other mass-loss phases, is not well understood).

- The radius decreases and the temperature increases, with the luminosity about constant.
- From this point onward it is *useful to consider the evolution of the core and the envelope separately*.

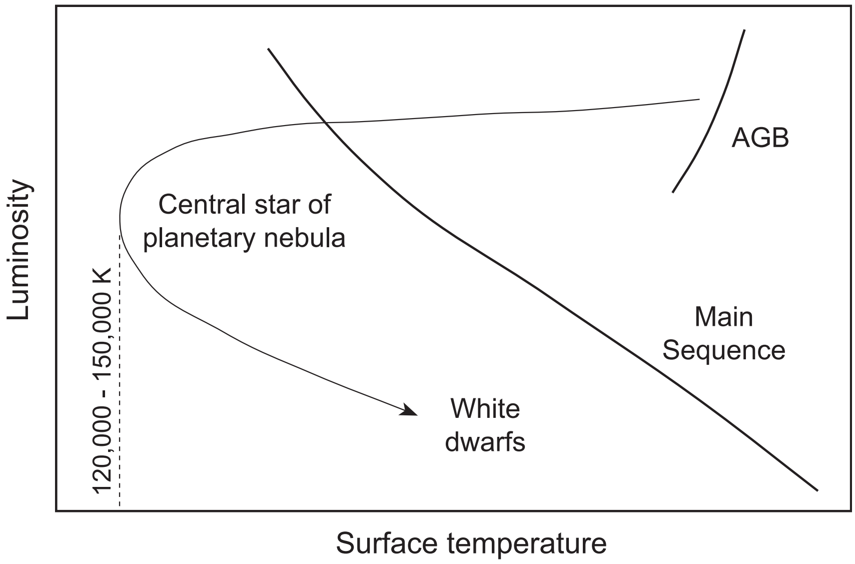


Figure 13.14: Evolution after the asymptotic giant branch.

As the core compresses, it follows the approximate evolutionary track shown in Fig. 13.14.

- This takes it to much higher temperatures than for the normal HR diagram.
- It finally cools to the white dwarf region with attendant decrease in luminosity.
- This high temperature is a result of
  - retained thermal energy
  - gravitational compression

  since the core is no longer capable of producing energy by thermonuclear processes.

- The remnants of the ejected envelope recede from the star.
- When the temperature of the bare core reaches about 35,000 K
  - a *fast wind*, probably associated with radiation pressure from the hot core, *accelerates the last portion of the envelope* to leave,
  - this forms a *shock wave* that proceeds outward and defines the inner boundary of the emitted cloud.
- As the temperature of the central star climbs, the *spectrum is shifted far into the UV*.
- The resulting bath of high energy photons from the central star *ionizes the hydrogen* in the receding envelope.
- The resulting *recombination reactions between ions and electrons emit visible light* and account for the luminosity and the often beautiful colors associated with the planetary nebula.
- The core and the planetary nebula now proceed on their separate ways:
  - the *core cools slowly to a white dwarf*,
  - the *planetary nebula expands and grows fainter*, eventually dispersing into the interstellar medium.

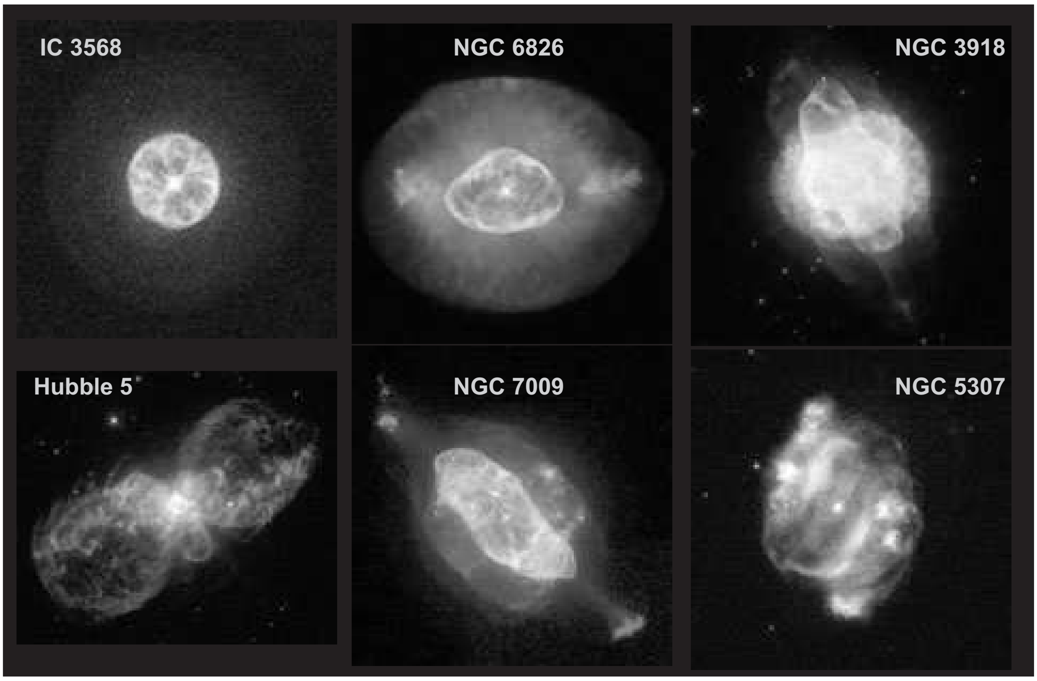


Figure 13.15: A variety of planetary nebulae imaged by the Hubble Space Telescope. Such observations indicate that the ways in which dying AGB stars eject their envelopes can be quite complex.

13.10 Stellar Dredging Operations

Observations indicate that *red giant stars exhibit isotopic abundances in their surfaces and winds* that could only have been produced by *nuclear burning in core and shell sources*.

- Post main-sequence evolution in the red giant region involves various *episodes of deep convection*.
- Thus it is logical to assume that the observed nuclear-processed material is brought to the surface by such *deep convective mixing*.
- This mechanism of transporting the products of nuclear burning and processing to the surface by deep convection is termed a *dredge-up*.

Three dredge-up episodes have been identified in post main-sequence evolution:

1. *First dredge-up* is thought to occur as the star develops deep convection driven by the hot hydrogen shell source *prior to triple-α ignition on the RGB*.
2. *Second dredge-up* can occur early in AGB evolution for intermediate-mass main sequence stars as a result of convective gradients *generated by the narrowing helium shell source*.
3. *Third dredge-up* appears to be necessary to understand surface abundances for many evolved AGB stars. It is thought to be associated in a complex way with
   - *thermal pulses in AGB evolution*, through deep convection that
   - extends at least periodically into the *region between the H and He shell sources*

Although these dredge-up episodes are only partially understood, they are key to explaining observations like

- *carbon stars* (stars with a greater abundance of C than O in their surfaces) and
- abundance of *interstellar carbon dust grains*.

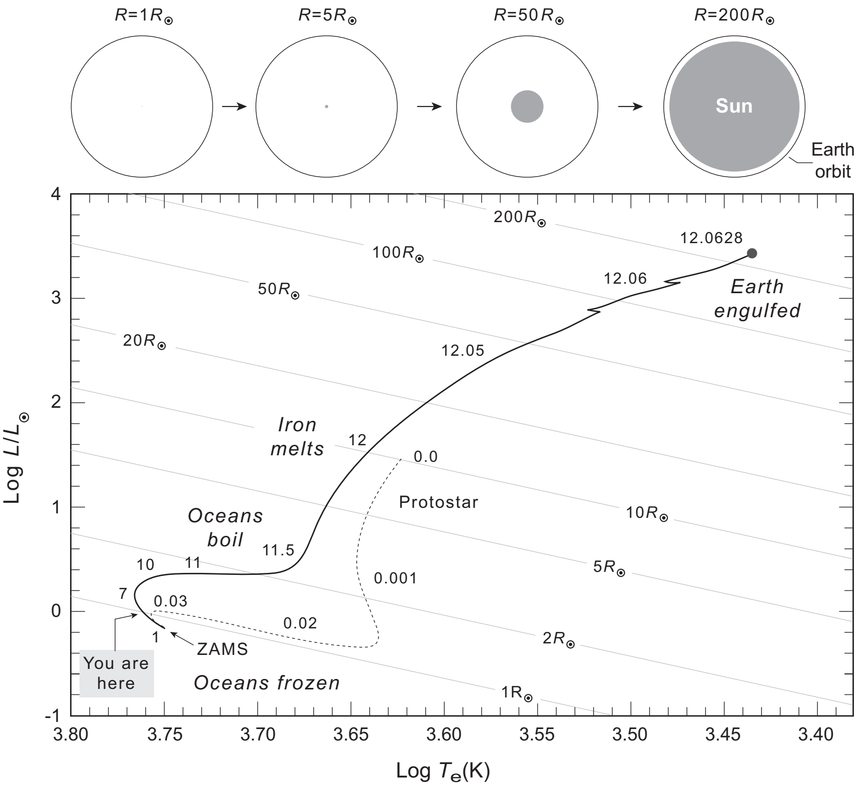


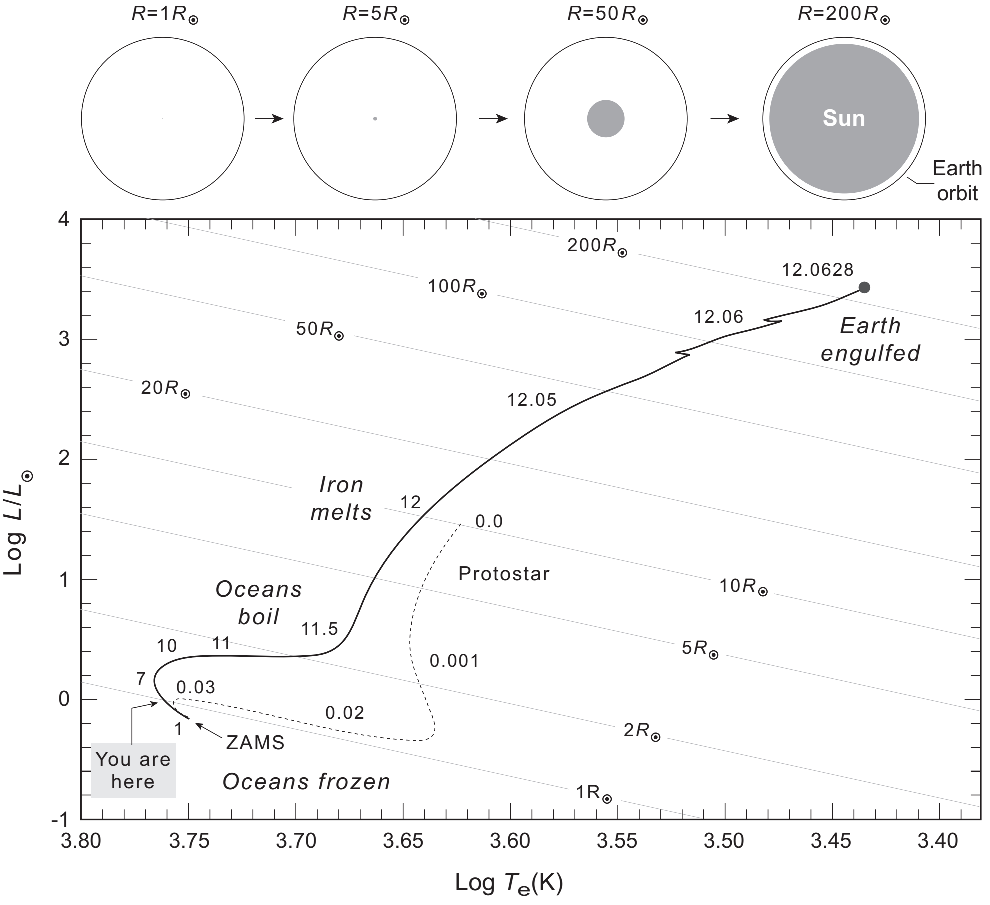
Figure 13.16: Evolution of the Sun showing implications for Earth. Times in units of 10^9 years are shown beside the curve. The parallel diagonal lines join points of constant stellar radius. Protostar evolution is indicated by the dotted curve, beginning from when the protostar has collapsed to a radius 10 times that of the present Sun. The top panel shows Earth's orbit and the Sun drawn to scale at various stages of the evolution.

13.11 The Sun's Red Giant Evolution

The Sun will

- *evolve into a red giant* and
- shed its envelope to become a *white dwarf*,

with consequences for Earth.



- The Sun has *5 billion years left on the main sequence*.
- The above figure illustrates a simulation for expansion of the Sun in the beginning of its red-giant phase.
- The top panel shows *the Sun and Earth's orbit drawn to scale* at various stages of the evolution.
- Presently the *Earth's orbital radius is 214 times the radius of the Sun*, so the Sun is largely invisible on this scale.
- In this simulation the *Sun expands to the size of Earth's present orbit* 12.0628×10^9 yr after the time marked zero.

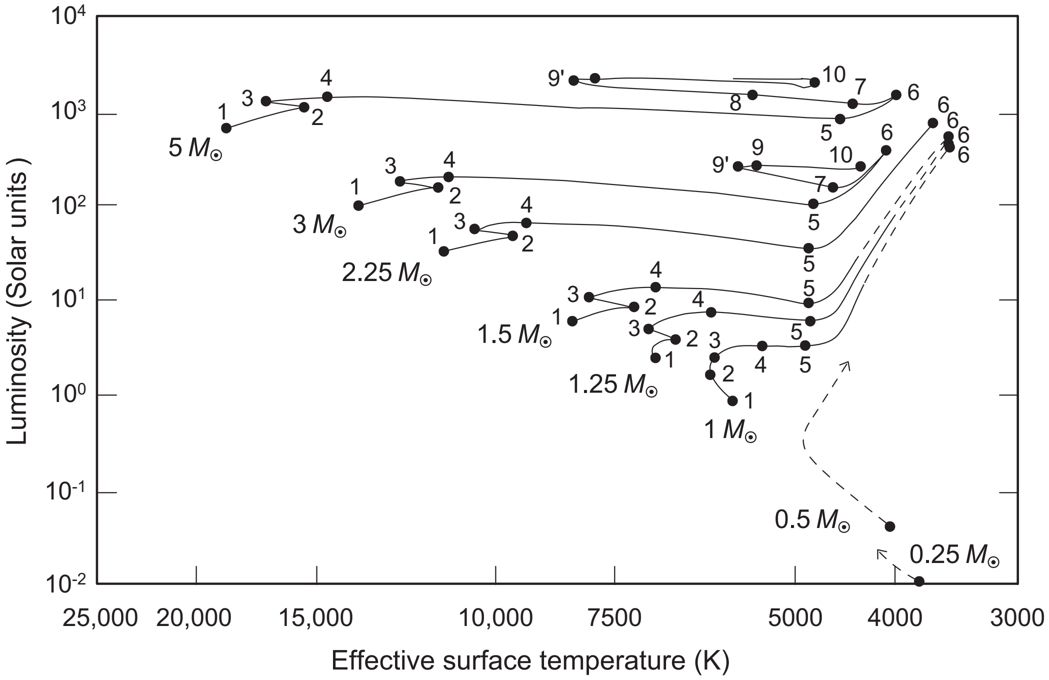


Figure 13.17: Evolution off the main sequence for stars of 5 $M_\odot$ or less.

13.12 Overview for Lower-Mass Stars

An overview of evolution after leaving the main sequence for various stars in the 0.25–5 $M_\odot$ range is given in Fig. 13.17.

- All but the lightest evolve into the red giant region but they exhibit mass-dependent differences:
  - *Evolution is faster for the heavier stars*, with possible looping and switchbacks.
  - *For higher mass* the post main-sequence HR motion is increasingly *horizontal and to the right*.
  - For *lower mass* the ascent is *highly-vertical*.

Chapter 14

Evolution of Higher-Mass Stars

In this chapter we address the evolution of high-mass stars, which will be defined as a ZAMS mass $M \gtrsim 8M_\odot$. There are some important issues that are unique to high-mass stars:

- The same burning stages as for lower-mass stars are encountered, but *additional advanced burning stages of heavier fuels are initiated* that are not accessible to lower-mass stars.
- The evolution through all stages occurs *more rapidly and at greater luminosity* than for less-massive stars.
- Nucleosynthesis occurring in evolution after the main sequence produces *heavier and more varied elements* than those synthesized for less-massive stars.
- The role of neutrino emission becomes increasingly pronounced in more advanced burning stages, with *core-cooling dominated by neutrinos for carbon burning and beyond.*

- The luminosity on the main sequence and after is often *close to the Eddington limit* and remains relativity constant after the main sequence.
- Thus the *evolution after the main sequence for massive stars is very horizontal on the HR diagram.*
- *Mass loss by strong stellar winds can be significant,* even on the main sequence.
- The central temperatures are high and the *core electrons typically remain nondegenerate until the latest burning stages,* despite the high density.
- The iron core formed in the last stages of main-sequence evolution for massive stars is supported by electron degeneracy pressure and is inherently unstable if it grows beyond a critical mass.
- This implies that the *endpoint of stellar evolution will be fundamentally different for a massive star* relative to that for an low-mass star.

Each of these issues will be addressed in this chapter or in the discussion of core collapse supernova explosions in Ch. 20. It is useful to begin with the consequences of advanced burning stages for massive stars.

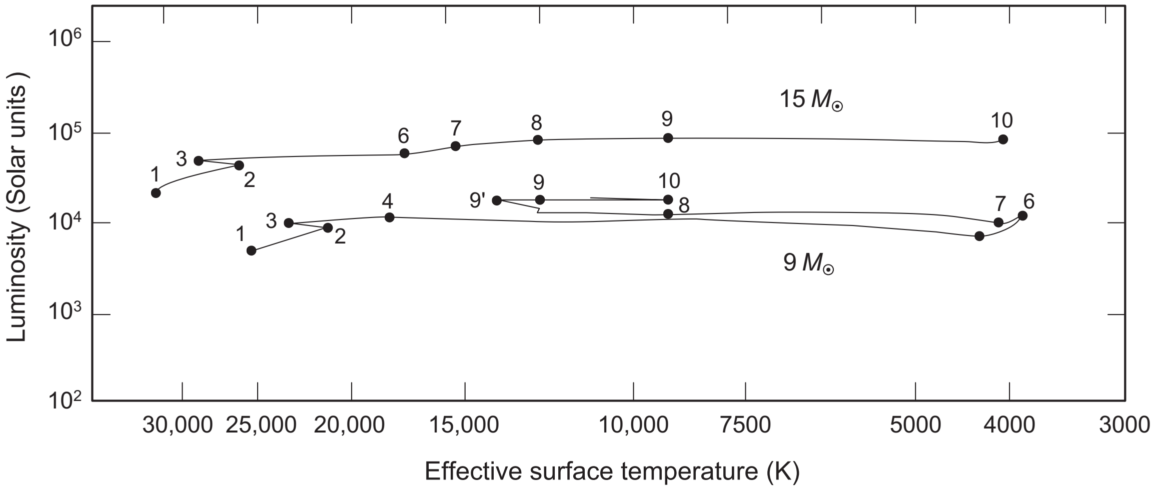


Figure 14.1: Evolution off the main sequence for high-mass stars.

14.1 Advanced Burning Stages in Massive Stars

An overview of evolution for two stars in the $M > 8M_\odot$ range after leaving the main sequence is given in Fig. 14.1.

- The rate of evolution after the main sequence for these stars is extremely rapid.
- For example, from the 9 $M_\odot$ star reaches point 10 in Fig. 14.1 only about *5 million years* after leaving the main sequence.

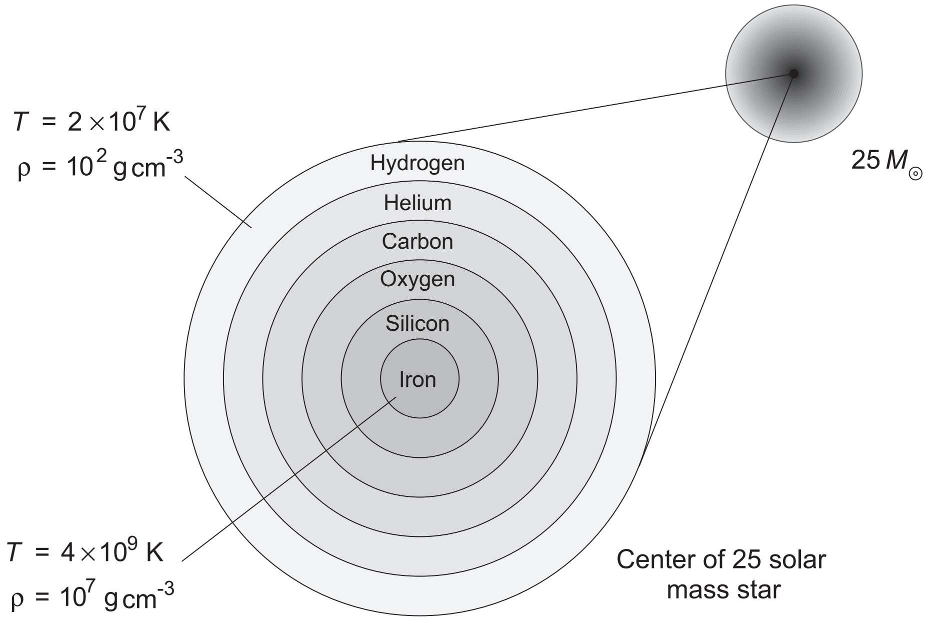
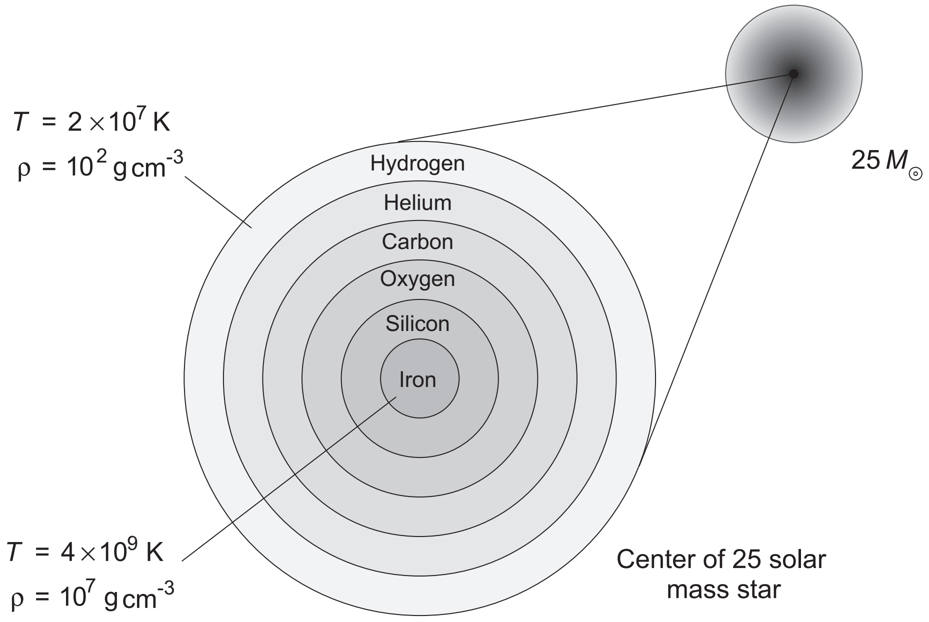


Figure 14.2: Central region of a 25 solar mass star late in its life. This central region is only a few thousand kilometers in radius and lies at the center of a supergiant star.

Because of sequential advanced burning stages, massive stars build up the layered structure depicted schematically in Fig. 14.2.

- If the star has a mass $M \gtrsim 8M_\odot$, successively heavier fuels can be burned as the star compress and heats up, until *an iron core is formed in the center of the star.*
- The iron core *cannot produce energy by fusion*.
- The iron core is supported by *electron degeneracy pressure.*



- The silicon layer surrounding the iron undergoes reactions producing more iron and *the iron core grows more massive.*

- Beyond a critical mass of about $1.2\,M_\odot$ the core becomes *gravitationally unstable and collapses.*

- This collapse will be described in some detail later.

- Here we will concentrate on describing the evolution of high-mass stars prior to encountering the gravitational instability of the iron core.

14.1.1 Envelope Loss from Massive Young Stars

As we discuss later, there is strong observational evidence that very massive stars eject large amounts of material from their envelopes early in their lives. It is thought that

- Radiation pressure and
- pulsational instabilities

play a leading role in these mass-loss processes.

Massive stars may go through early stages where they expel large portions of their envelopes into space at velocities as large as $1000\,\mathrm{km\,s^{-1}}$.

- In such stars, the timescale for mass loss

  $$\tau_{\mathrm{loss}} \sim \frac{M}{\dot{M}},$$

  where M is the mass and $\dot{M}$ the rate of mass loss, may be shorter than their main sequence timescales.

- One class of stars exhibiting large mass loss is that of *Wolf–Rayet stars,* which are characterized by

  1. large luminosity,
  2. envelopes strongly depleted in hydrogen, and
  3. high rates of mass loss.

- Most Wolf–Rayet stars have masses of $5-10 M_\odot$.

- They are thought to be the remains of stars initially more massive than $30 M_\odot$ that have

  - ejected all or most of their outer envelope,
  - exposing the hot helium core.

- The envelopes of Wolf–Rayet stars typically contain *10% or less hydrogen by mass*, with individual stars exhibiting different levels of envelope stripping.

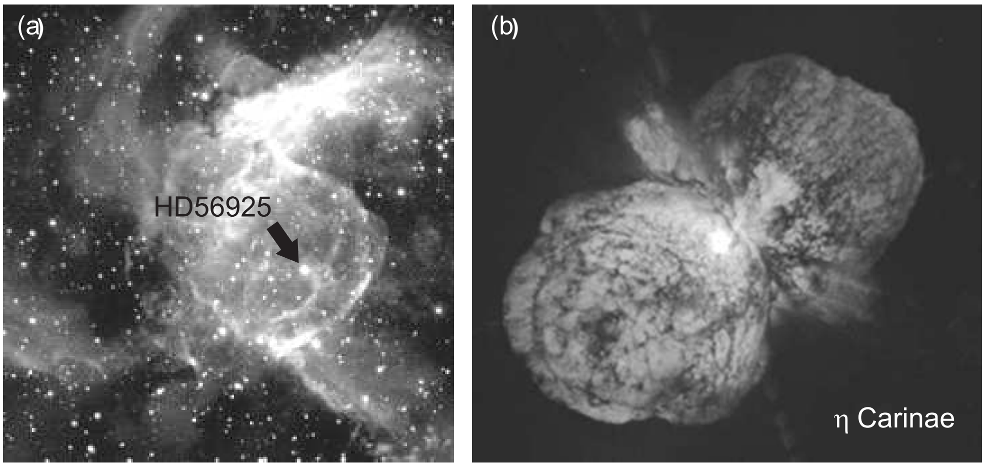


Figure 14.3: (a) Wolf–Rayet star HD56925 surrounded by remnants of its former envelope. (b) η Carinae, surrounded by ejected material.

- Figure 14.3a shows the nebula N2359, a wind-blown shell of gas that has been expelled from the *Wolf–Rayet star HD56925* (marked by the arrow.)

- The nebula contains shock waves generated by interaction of the wind and interstellar medium, and is glowing from excitation of expelled material.

- Figure 14.3b shows an extreme example of mass loss: the supermassive, highly unstable star, η *Carinae*.

- Elemental abundances in the nebula around η Carinae are consistent with this being the supergiant phase of a $120\ M_\odot$ star that has evolved with very large mass loss on the main sequence and afterwards.

- (η Carinae may be a binary or triple star, which complicates details but not the essence of the interpretation.)

14.2 Neutrino Cooling of Massive Stars

In advanced burning stages *neutrino cooling* is important.

- The conditions in stars leading to these burning stages often involve extreme energy-production rates in regions deep within stars having high photon opacity.
- Then neither radiative nor convective transport can remove the energy fast enough to maintain equilibrium.
- But the high-temperature, high-density environment is at the same time conducive to neutrino production and the material is still transparent to neutrinos that are produced.
- Hence the very stability of stars undergoing advanced burning depends fundamentally on neutrino cooling.
- This in turn implies that the properties of late stellar evolution and the types of remnants that result are bound up inextricably with neutrino cooling of the star.
- Neutrino cooling assumes particular importance for high-mass stars, which can access all the advanced burning stages that have been discussed.

From carbon burning and beyond the dominant mode of cooling in stellar evolution becomes neutrino emission.

14.2.1 Local and Non-Local Cooling

Below temperatures of about 5×10^8 K stars are cooled dominantly by radiation and convection.

- The net rate of heat removal *depends on temperature gradients*.
- Thus the cooling at a given point in the star is *nonlocal,* in that it depends on conditions in the surrounding region.

In contrast, neutrino cooling is highly *local:*

- The energy carried by a neutrino produced at a point
  - is removed from the star at nearly the speed of light ,
  - with little probability to interact with any of the rest of the star.
- Thus neutrino cooling depends only on the *conditions at the point of production* and not on spatial derivatives evaluated at that point.

14.2.2 Neutrino Cooling and the Pace of Stellar Evolution

Neutrino emission begins to dominate the energy-removal budget in stars when temperatures exceed about 10^9 K and densities are sufficiently low that the electrons are not too degenerate.

- However, it should be noted that "neutrino cooling"
  - is apt when applied to white dwarfs or neutron stars, but
  - is something of a misnomer for stars undergoing thermonuclear burning in hydrostatic equilibrium.
- Instead of cooling the star, the rapid energy loss from neutrino emission *stimulates increased thermonuclear rates* that are required to keep the star in equilibrium.

Thus, neutrino "cooling" actually *accelerates burning and the pace of stellar evolution* for massive stars.

14.3 Massive Population III stars

An interesting and exotic aspect of massive star evolution concerns the first generation of stars that formed in the Universe *(Population III)*.

- Observations suggest that the first stars began forming *several hundred million years after the big bang.*
- These stars would have been
  - *hydrogen and helium stars* with
  - *negligible metals,*

  since they formed from material produced by the big bang.
- Simulations indicate that because of their low metal content *these stars likely were very massive,* with 100–1000 $M_\odot$ being common.
- Because of their large mass,
  - these stars would have *evolved quickly* and
  - most would have exploded as *pair-instability supernovae* within several million years of their birth.
- These explosions *seeded the Universe with heavier elements* up to iron.

At the *recombination transition* in the early Universe, which occurred at a redshift $z \sim 1100$ (some 380,000 years after the big bang),

- Electrons combined with protons to make neutral hydrogen and *the Universe became transparent.*
- There were no stars yet, so the ensuing period until stars formed is called the *dark ages.*
- From redshift $z \sim 20$ to $z \sim 6$ (roughly from 500 million years to almost a billion years after the big bang) the neutral hydrogen was reionized in the *reionization transition.*
- It is widely believed that *Pop III stars were responsible for this reionization* of the Universe.
- Spectra of high-redshift quasars indicate that there were *heavy elements present during reionization.*
- These could have come only from stars, and because of their large masses stars in this first generation would have been hot and would have bathed their neighborhoods with ionizing UV radiation.
- No conclusive observational evidence exists for Pop III stars.
- Some candidates have been proposed for star clusters found in faint galaxies at large redshift ($z \geq 6$), but the observations are difficult and thus conclusions are necessarily qualified.

14.4 Evolutionary Endpoints for Massive Stars

Stars having $M \lesssim 8M_\odot$ when they leave the main sequence are all thought to end their lives with

- their cores evolving to some form of *white dwarf* (helium, carbon–oxygen, or neon-magnesium), and
- their envelopes ejected as *planetary nebulae.*

In contrast, the most massive stars appear ordained to one of three qualitatively different fates more dramatic than becoming white dwarfs, with all three initiated by *gravitational collapse of the star's core:*

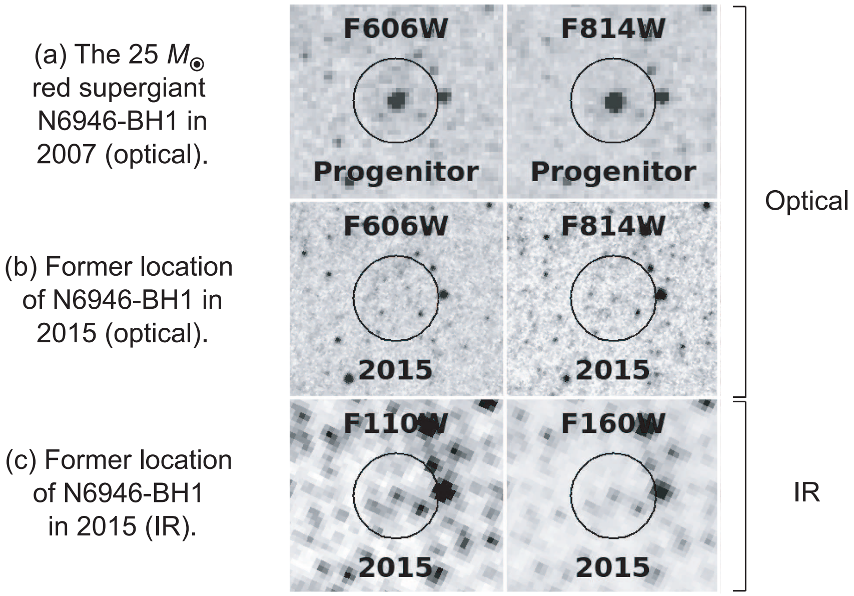
- The majority of stars having $M \gtrsim 8M_\odot$ will eject the outer layers of the star violently in a *core collapse supernova,* with the central regions crushed gravitationally into a *neutron star* that is stabilized by neutron degeneracy pressure.
- For some core collapse events the mass of the gravitationally-collapsed central region will be too large for neutron degeneracy pressure to halt the infall and the star will collapse instead to a *black hole.*
- For the special case of very massive stars ($M \sim 130 - 250\,M_\odot$) and low metallicities, the star can destroy itself in a *pair-instability supernova,* which leaves behind no compact remnant. This was probably the fate of many Pop III stars.

14.4.1 Observational and Theoretical Characteristics

The neutron star and black hole endpoints for core collapse likely have different observational characteristics.

- For the neutron star outcome
  1. gravitational waves and a burst of neutrinos will be emitted from the supernova explosion,
  2. the ejected outer layers will produce an expanding supernova remnant, and
  3. the neutron star will cool primarily by neutrino emission, perhaps manifesting as a pulsar.
- For the black hole outcome,
  1. a direct collapse to a black hole is expected to produce bursts of neutrinos, gravitational waves, and possibly γ-ray bursts, but
  2. little in the way of traditional supernova remnants may be ejected.

> Hence it is possible that some or all core collapse events leading to black hole formation are “dark”, with only the emission of gravitational waves, neutrinos, and possibly gamma-ray bursts, but no traditional astronomy observations, to mark their passage.


(a) The 25 $M_\odot$ red supergiant N6946-BH1 in 2007 (optical).

(b) Former location of N6946-BH1 in 2015 (optical).

(c) Former location of N6946-BH1 in 2015 (IR).

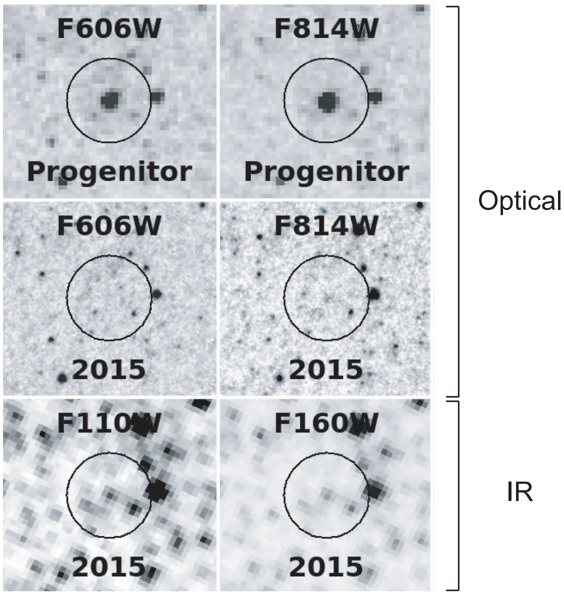
Figure 14.4: Evidence for a failed supernova. HST optical and IR images of the region surrounding the 25 $M_\odot$ red supergiant N6946-BH1. (a) In these optical images from July, 2007, N6946-BH1 is the spot at the center of the circles, which have radii of 1 arcsec. (b) in optical images of the same region from October, 2015, N6946-BH1 has disappeared. (c) In 2015 very faint IR emission was observed consistent with the former position of N6946-BH1.

14.4.2 Black Holes from Failed Supernovae?

Can massive stars collapse directly to black holes, without ejection of a remnant and without a large increase in optical luminosity? Such *failed supernovae* would produce

- a black hole,
- gravitational waves, and
- neutrinos,

but few other characteristics of core collapse supernovae.



(a) The 25 $M_\odot$ red supergiant N6946-BH1 in 2007 (optical).

(b) Former location of N6946-BH1 in 2015 (optical).

(c) Former location of N6946-BH1 in 2015 (IR).

- In 2007 the $25\,M_\odot$ red supergiant N6946-BH1 appears as a dark spot in HST optical images (circled in above figure).
- In 2009 this star brightening to $L \geq 10^6 L_\odot$ but then faded to less than pre-outburst luminosity over a few months.
- Images from 2015 [Fig. (b)] indicate that N6946-BH1 has disappeared in the optical, but Fig. (c) indicates faint IR emission at the former location of N6946-BH1.
- The luminosity of N6946-BH1 in 2017 was much less than the progenitor, suggesting that the star is no more.

These observations are consistent with the $25\,M_\odot$ supergiant undergoing a *failed supernova* and collapsing *directly to a black hole,* with faint residual IR from weak accretion on the black hole.

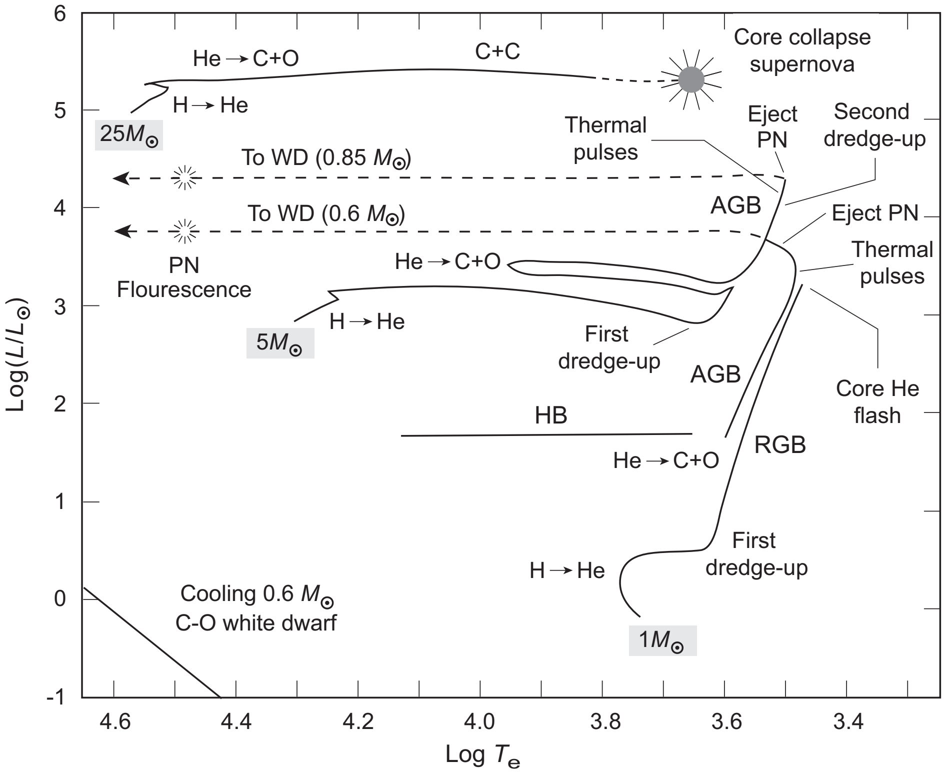


Figure 14.5: Summary of late stellar evolution: HB (horizontal branch), RGB (red giant branch), AGB (asymptotic giant branch), PN (planetary nebula), and WD (white dwarf). Reactions in burning stages are indicated.

14.5 Summary: Evolution after the Main Dequence

A summary of late stellar evolution is shown in Fig. 14.5.

- *Low-mass stars* evolve slowly to white dwarfs, with emission of the envelope as a planetary nebula.
- *High-mass stars* evolve quickly to catastrophic core collapse, leaving behind a neutron star or black hole.

This is yet another installment in the ongoing saga: for stars, *mass is destiny.*

14.6 Stellar Lifecycles

We conclude this chapter by noting that stellar evolution leads to extensive recycling of stellar material.

- Each star ties up a certain amount of mass at its birth.
- As the star evolves, some of that mass is returned to the interstellar medium to participate in future star formation by winds and explosions.
- Some becomes locked forever in white dwarfs, neutron stars, and black holes, assuming that white dwarfs and neutron stars don't undergo interactions with other objects after their formation.

The birth, evolution, and death of successive generations of stars has three general consequences for a galaxy:

1. The gas available to make stars decreases as more of it becomes locked in white dwarfs, neutron stars, black holes, brown dwarfs, and very low mass stars .
2. Over time the *luminosity declines* and the *light reddens* as
   - massive, bright stars die more quickly and
   - the population is increasingly dominated by less-massive, long-lived, fainter stars.
3. The gas in the galaxy becomes *enriched in metals* as nuclear-processed material is returned from stars to the interstellar medium by winds and explosions.

Thus, successive generations of stars typically have higher metallicities. However,

- The metal content does not increase uniformly with time.
- Example: From metallicities of stars with different ages in the Milky Way it may be estimated that the mass fraction of heavy elements Z increased by much more early in the history of the galaxy than it has more recently.
- The contribution to metallicity also is not uniform with star mass.
- Massive stars are rare but they are the primary source of metallicity increase because they eject large amounts of processed mass as winds and explosions on a relatively short timescale.

> - The fraction and composition of stellar material returned to the interstellar medium depends strongly on the mass of a star.
> - Thus the *initial mass function* discussed earlier is important in understanding the recycling of stellar material.

Chapter 15

Stellar Pulsations and Variability

One commonplace of modern astronomy that would have been highly perplexing for ancient astronomers is that many stars vary their light output by detectable amounts over time.

- In some cases these variations are asynchronous and in others they are highly periodic.
- They may be so small as to require precise instruments to detect them, or sufficiently large that they are easily visible to the naked eye.

These *variable stars* may be loosely classified into three categories.

1. *Eclipsing binaries* are binary stars in which the total light output of the system is altered by geometrical eclipses of one star by the other. If the binary system is too far away to resolve the two components, this will appear to be a single star with periodic variation in light output.

2. *Eruptive and exploding variables* are stars that suddenly increase light output and often eject mass because of a rapid and violent disruption or partial disruption of the star. Novae and supernovae are dramatic examples in this category.

3. *Pulsating variable stars* appear to undergo (possibly complex) pulsations that alter the light output in periodic or irregular fashion, without disrupting significantly the overall structure of the star, Well-known examples of this category are Cepheid variables and RR-Lyrae stars.

In this chapter we examine in more depth this latter category and the reasons that some stars become unstable against pulsations for certain periods of their lives.

Table 15.1: Pulsating variable stars

Variable type	Period	Population	Mode[†]
Long-period variables	100–700 d	I, II	R
Classical Cepheids	1–50 d	I	R
Type-II Cepheids	2–45 d	II	R
RR Lyrae stars	1.5–24 hr	II	R
δ Scuti stars	1–3 hr	I	R, NR
β Cephei stars	3–7 hr	I	R, NR
ZZ Ceti stars	100–1000 s	I	NR

[†] R = Radial; NR = Non-radial

15.1 The Instability Strip

Some common classes of pulsating variable stars and their characteristics are given in Table 15.1.

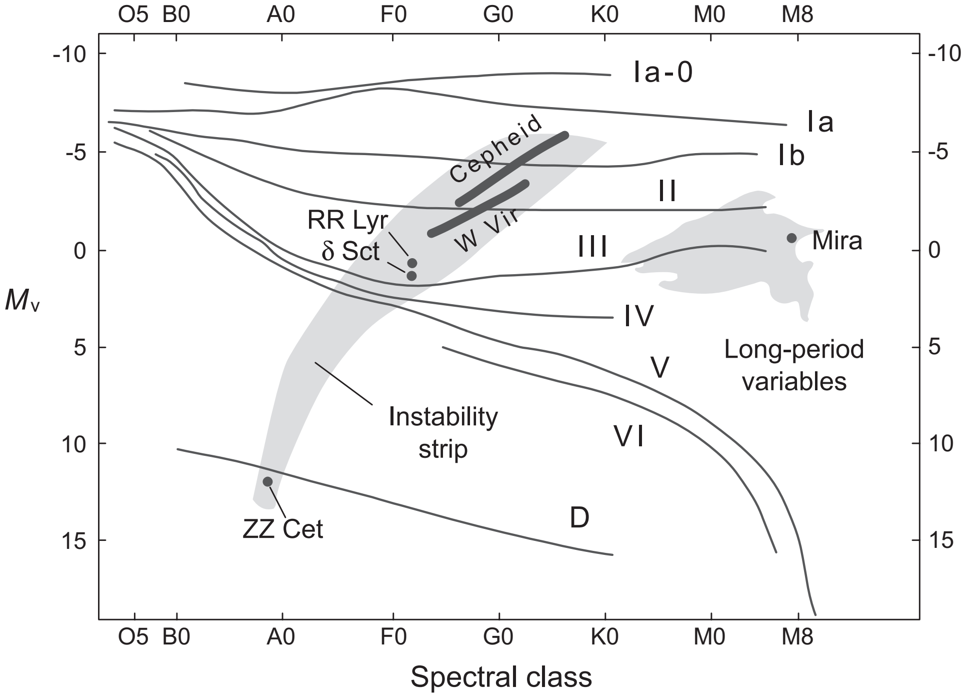


Figure 15.1: The instability strip and the region of long-period variables in the HR diagram. With the exception of the long-period variables, most variable stars are found within the instability strip.

Pulsating variables are found in specific regions of the HR diagram, as illustrated in Fig. 15.1.

- There we see that many types of pulsating variables are confined to a narrow, rather vertical, strip in the HR diagram called the *instability strip.*
- This suggests that there is a fundamental mechanism
  - operating in various stars of different luminosity, but
  - over a narrow range of surface temperatures,

  that leads to pulsational instability.

15.2 Adiabatic Radial Pulsations

At the simplest level we may examine stellar pulsation in terms of oscillations within the body of the star that are *adiabatic and linear in the displacement,* and that maintain spherical symmetry for the star.

- Such an analysis has much in common with the study of small-amplitude vibrations in other physical systems:
  - The pulsations are treated as as free radial vibrations.
  - Gas compression plays the role of a spring constant.
- One finds that stars disturbed slightly from spherical hydrostatic equilibrium exhibit discrete vibrational frequencies that are called *radial acoustic modes*.

15.2.1 Radial Acoustic Modes

It is convenient to discuss stellar pulsation in terms of Lagrangian coordinates, where $m(r)$ is the independent variable and corresponds to the mass contained within a radius r.

- Then if we expand the pressure, radial coordinate, and density as time-dependent oscillations around the equilibrium values (which are denoted by a subscript zero),

$$\begin{aligned} P(m,t) &= P_0(m)\left(1+\delta P(m)e^{i\omega t}\right) \\ r(m,t) &= r_0(m)\left(1+\delta r(m)e^{i\omega t}\right) \\ \rho(m,t) &= \rho_0(m)\left(1+\delta\rho(m)e^{i\omega t}\right), \end{aligned}$$

  where the radial displacement $\delta r(m)$ is described by

$$\frac{d^2(\delta r)}{dr_0^2} + \left(\frac{4}{r_0} - \frac{\rho_0 g_0}{P_0}\right)\frac{d(\delta r)}{dr_0} + \frac{\rho_0}{\Gamma_1 P_0}\left[\omega^2 + (4-3\Gamma_1)\frac{g_0}{r_0}\right]\delta r = 0,$$

where Γ_1 is an adiabatic exponent, $g_0 \equiv Gm/r_0^2$, and ω is the adiabatic oscillation frequency.

- This equation must be solved with two boundary conditions, one at the center of the star and one at the surface.
  - At the center one requires $d(\delta r)/dr_0 = 0$.
  - The simplest physically reasonable surface boundary condition is to require $\delta PP_0 = 0$.

Most intrinsically variable stars are pulsing in radial acoustic modes, which correspond to standing waves within the star.

- The *fundamental mode* has no nodes (points of zero motion) between the center and surface, implying that the stellar matter involved in the vibration all moves in the same direction at a given time.
- The *first overtone* has one node between the center and the surface, so the matter moves in one direction outside this node and in the opposite direction inside this node at a given phase of the pulsation.
- Likewise, higher overtones with additional nodes and more complex motion may be defined.
- Just as for musical instruments and other acoustically vibrating systems, a star may exhibit several modes of oscillation at once.
- The physical motion of the gas in radial stellar pulsations is largest in the fundamental mode and is considerably smaller in the first overtone.
- In higher overtones the motion of the gas in an oscillation cycle is even smaller.

Pulsating variables appear to be oscillating primarily in the *fundamental mode and/or the first overtone*.

It is thought that

- most Classical and Type II Cepheids oscillate in the fundamental mode, while
- RR Lyrae stars oscillate in either the fundamental mode or first overtone (or both).

For long-period red variables the evidence is less conclusive and they may pulsate in either the fundamental or first overtone modes.

15.2.2 Pulsations in Realistic Stars

Pulsations in realistic stars are more complicated than the linear adiabatic analysis discussed in the preceding paragraphs would indicate.

- For example both the
  - *rate of energy production* and
  - the *rate of internal energy transport* may be modified by pulsations,

  so we may expect that they are *not completely adiabatic* and must examine deviations from adiabaticity.
- In particular, we must ask the question:*what energy input sustains the pulsation modes of a pulsating variable star?*

Eddington first examined systematically the idea that stellar pulsations are free radial oscillations, but realized that dissipation processes in the gas would damp out such oscillations quickly.

- Example: pulsations of Cepheid variables should be damped on a timescale of order 10^4 years without some mechanism to amplify and sustain the oscillations.
- Thus the steady, long-term pulsing of a Cepheid variable
  - cannot be due to a one-time excitation of eigenmodes and
  - cannot be adiabatic.

Eddington proposed that pulsating variable stars are a form of *heat engine* continuously transforming thermal energy into the mechanical energy of the pulsation.

On the other hand, it will turn out that in realistic stars the pulsation may often be approximated as *nearly adiabatic:*

- Instabilities grow on a timescale that is *long relative to the time for one pulsation.*
- Over one acoustic oscillation cycle (which is essentially a *hydrodynamic timescale*), the amount of heat exchanged is typically small.
- This is because energy transfer occurs on a *Kelvin–Helmholtz timescale*, which is much longer than the hydrodynamic timescale.
- Therefore after a single acoustic cycle the star returns almost—*but not quite*—to the initial state.
- The *"not quite"* measures the lack of reversibility and therefore the non-adiabaticity of the process.

With this as introduction, let us now consider the role of non-adiabatic effects in sustaining stellar pulsation.

15.3 Non-Adiabatic Radial Pulsations

For each layer of the star a net amount of work is done during a pulsation cycle that must be equal to the difference of the heat flowing into that layer and that flowing out.

- If the oscillation is to be self-sustaining for a single layer, we must have a mechanism whereby
  - heat enters the layer at high temperature and
  - heat leaves the layer at low temperature.
- If layers driving the pulsation absorb energy near the time of maximum compression, the oscillations will be amplified because the time of maximum pressure in the layer will occur after maximum compression.

> This is similar to the reason that it is optimal to fire the spark plug near the *end* of the compression stroke in an internal combustion engine.

- A sustained oscillation for a significant part of the star then requires that a set of different layers have some level of phase coherence in these driven oscillations.

Let us now justify these assertions using basic ideas from thermodynamics.

15.3.1 Thermodynamics of Sustained Pulsation

Many features required to sustain stellar pulsation follow from the 1st and 2nd laws of thermodynamics.

- Let's work in Lagrangian coordinates and assume the gas to be almost but not quite adiabatic.
- Consider a mass zone. By the 1st law, for a pulsation cycle the change in heat Q for a mass zone is a sum of contributions from changes in internal energy U and work W done on its surroundings during the pulsation,

$$dQ = dU + dW.$$

- After a complete oscillation cycle we assume that the internal energy U returns to its original value so that the work done over the cycle is entirely contributed by the change in Q,

$$W = \oint dQ.$$

- To drive oscillations, the gas must do positive work on its surroundings (absorb heat). However, we assume the system to be nearly adiabatic, so the gas returns essentially to its original state after one cycle.
- Therefore, *in zero order* there is no change in entropy

$$\oint dS = \oint \frac{dQ}{T} = 0.$$

- Now suppose that during the cycle we perturb the system by a small periodic variation in the temperature T of the form
$$T = T_0 + \Delta T(t),$$
where $\Delta T = 0$ at the beginning and end of the cycle.

- Then
$$\oint dS = \oint \frac{dQ}{T} = 0 \quad \rightarrow \quad \oint \frac{dQ(t)}{T_0 + \Delta T(t)} = 0.$$

- Assuming the variation in T to be small, we expand the denominator of the integrand to first order and obtain
$$\oint dQ(t)\,(T_0 - \Delta T(t)) = 0,$$
or upon rearrangement,
$$\oint dQ(t) = \oint \frac{\Delta T(t)}{T_0} dQ(t).$$

- Then for the work done in one pulsation cycle
$$\left.\begin{array}{r} W = \oint dQ \\ \oint dQ(t) = \oint \dfrac{\Delta T(t)}{T_0} dQ(t) \end{array}\right\} \quad \rightarrow \quad W = \oint \frac{\Delta T}{T_0} dQ.$$

- For the cyclic integral

  $$W = \oint \frac{\Delta T}{T_0} dQ$$

  to give a net positive value (so that the mass zone does work on its surroundings over one cycle and can therefore drive an oscillation), we see that generally

  > ΔT and dQ must have the *same sign* over a major part of the cycle.

- That is, heat must be

  - absorbed ($dQ > 0$) when the temperature is increasing in the cycle ($\Delta T > 0$), and
  - released ($dQ < 0$) when temperature is decreasing in the cycle ($\Delta T < 0$).

The preceding discussion has concentrated on the behavior of a single mass zone.

- Oscillation of the entire star means that some zones may do positive work and other zones may do negative work within a pulsation cycle.
- Thus, the condition for amplifying and sustaining oscillation of the entire star is that

  $$W = \sum_i W_i = \sum_i \oint \left(\frac{\Delta T}{T_0}\right)_i dQ_i > 0,$$

  where i labels the mass zones of the star.
- (Strictly this sum is an integral over the continuous mass coordinate, but in practical numerical simulations the zones are normally discretized.)

We must now ask whether there are situations in stars that allow this condition to be realized.

15.3.2 The Role of Radiative Opacity

One way to favor sustained oscillations is to arrange that the opacity *increases* as the gas in a layer is compressed.

- Then the radiative energy outflow can be trapped more efficiently by the layer (it begins to "dam up" the outward energy flow).
- This can push it and layers above it upward until
  - the layer becomes less opaque upon expansion,
  - the trapped energy is released,
  - the layer falls back to initiate another cycle.

If a sequence of layers one above the other behaves in this way, a sustained oscillation could be set up.

- Conversely, if compressing the layer increases T and thereby *decreases* κ, the layer allows heat to flow through it more easily than before the compression, implying that $dQ < 0$ while T is increasing.
- Likewise, decompression in the 2nd part of the pulsation cycle causes T to fall and κ to increase, which traps more heat and causes dQ to be positive.
- Thus heat flow works against the oscillation under these conditions and will tend to damp it.

15.3.3 Opacity and the κ-Mechanism

Normal stellar radiative opacities do not increase with compression of the gas.

- From the Kramers form

$$\kappa \sim \rho T^{-3.5}$$

the opacity κ is proportional to ρ and to $T^{-3.5}$.

- Compression of a layer increases both ρ and T.
- However, *the temperature dependence is much stronger than the density dependence* for κ.
- Thus a gas described by a Kramers opacity tends to experience a *decrease* in opacity under compression.

Hence a star exhibiting the usual opacity behavior has a *built-in damping mechanism* that stabilizes it against pulsations.

- This explains why most stars are *not* pulsating variables.

However, there is a *special situation* for which the opacity could be expected to increase with compression.

- If a layer contains *partially ionized gas,* a portion of the energy flowing into it can go into more ionization.
- The energy absorbed into internal electronic excitations is not available to increase temperature.
- Thus, if there is sufficient ionization during the compression portion of the pulsation cycle,
  - the effect on the opacity of the smaller rise in T
  - can be more than offset by the increase in ρ and

  compression can *increase the opacity*.
- Conversely, electron–ion recombination in the decompression portion of the cycle can release energy and lead to a decreased opacity.
- Then, in partial ionization zones
  - a layer can absorb heat during compression when T is high and
  - release it during expansion when T is low,

  thereby setting the stage for a sustained oscillation.

This heat-engine mechanism for driving oscillations through ionization-dependent opacity effects is called the *κ-mechanism.*

15.3.4 Partial Ionization Zones and the Instability Strip

The κ-mechanism provides a possible way to drive stellar oscillations, but where do we expect the κ-mechanism to be able to operate?

- For most stars there are two significant zones of partial ionization, corresponding to the possible stages of ionization for hydrogen and helium:

  1. The *hydrogen ionization zone,* where
     - hydrogen is ionizing (H I → H II) and
     - helium is undergoing first ionization (He I → He II).

     This region is broad and typically has a temperature in the range 10,000-15,000 K.
  2. The *helium ionization zone,* where second ionization of helium (He II → He III) occurs, typically at a temperature around 40,000 K.

From the preceding discussion, we may expect one or both of these ionization zones to play a role in driving the pulsations of many variable stars.

Example: Detailed analysis indicates that

- For *classical Cepheids* (and most variables found in the instability strip)
  - the pulsation is caused by the κ-mechanism,
  - primarily by forcing of the fundamental mode in the helium ionization zone.
- On the other hand, the *long-period red variables* (large AGB stars like Mira) are thought to be driven by hydrogen ionization zones.

Temperature Boundaries for the Instability Strip

The radial location of hydrogen and helium ionization zones in stars of particular surface temperatures, and onset of convection near the surface for stars with surface temperatures that are too low, are determining factors in producing the instability strip.

- The physical radius for the hydrogen and helium partial ionization zones within a given star will depend strongly on the effective surface temperature of that star.
- For stars with higher temperatures, ionization zones will be near the surface and there will be insufficient mass in the partially-ionized layers to drive sustained oscillations.
- If the surface temperature is too low, convection in the outer layers will undermine the κ-mechanism (detailed simulations show that convection interferes with the trapping effect and thus damps stellar pulsations).
- This suggests an optimal range of surface temperatures for which

  1. the ionization zones are *deep enough to drive sustained oscillations* by coupling to the fundamental and overtones of the characteristic vibrational frequencies (→ higher-temperature end of the optimal range),
  2. but for which the *convection is not strong enough* to invalidate the mechanism (→ lower-temperature end of the optimal range).

Thus, pulsating variables should be found in localized regions of the HR diagram.

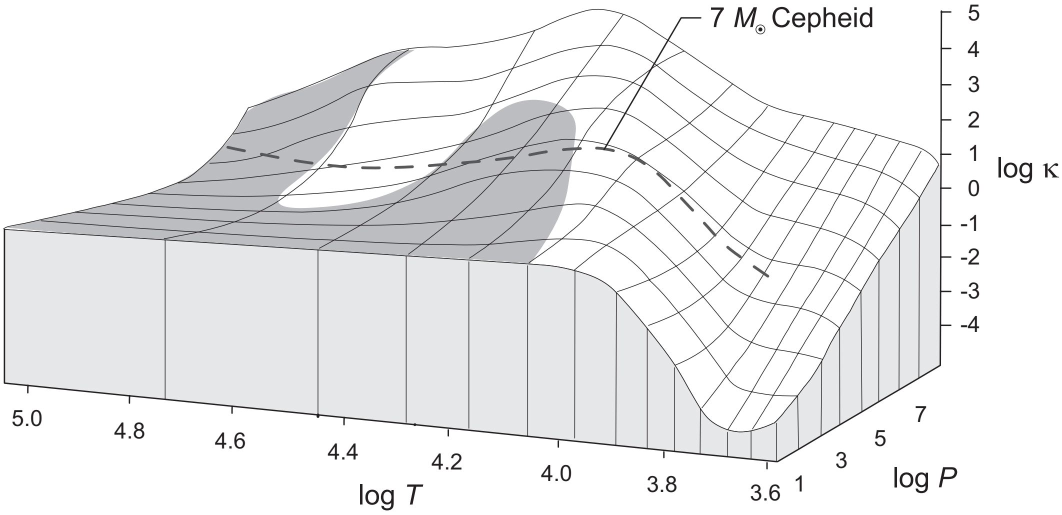
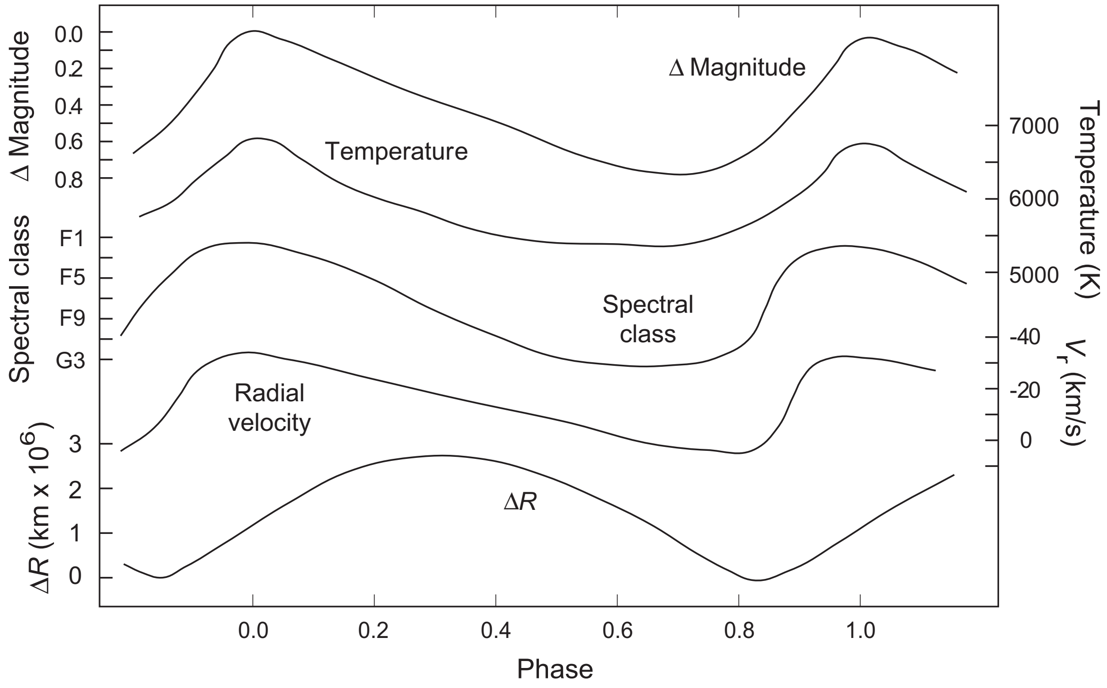


Figure 15.2: Rosseland mean opacity versus pressure and temperature.

15.3.5 Cepheid Variables and the Helium Ionization Zone

In Fig. 15.2 opacities expected for Cepheid variable stars are plotted as a function of temperature and pressure.

- Shaded regions correspond to conditions expected to damp oscillations and lighter regions represent conditions in which the opacity increases sufficiently with increased pressure to favor the κ-mechanism.
- The dashed line indicates the relationship between T and P expected for a 7 $M_\odot$ Cepheid variable.
- The helium ionization region crossed by the dashed line near $\log T = 4.6$ is thought to be the primary driver of classical Cepheid oscillations.

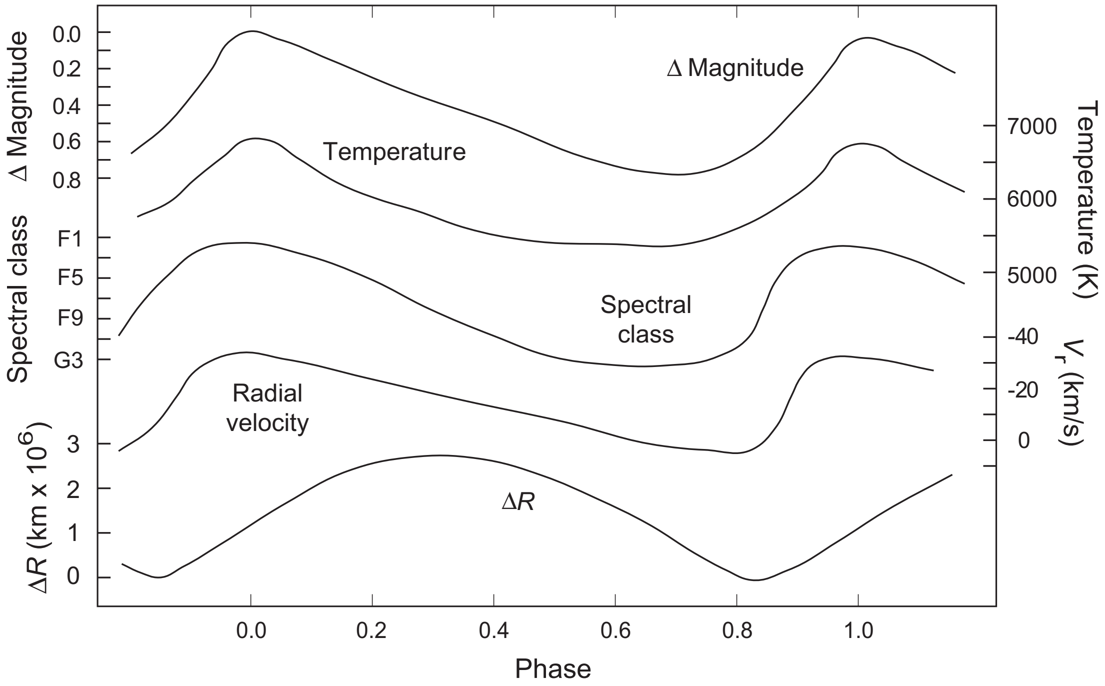


15.3.6 Cepheid Variables and the Hydrogen Ionization Zone

Helium ionization zones are primarily responsible for driving pulsations within the instability strip.

- However, the hydrogen ionization zones at lower temperature nearer the surface also play a (more subtle) role in the pulsation for stars like Cepheid variables and RR Lyrae stars.

- From the figure, maximum luminosity for a Cepheid is shifted systematically later relative to minimum radius in the pulsation *(maximum brightness for a Cepheid is correlated with maximum surface T*, not maximum radius).



- This is called the *phase lag,* and it is thought to be caused by oscillation of the hydrogen ionization zone toward and away from the surface.

- Simulations indicate that at minimum radius the luminosity at the base of the hydrogen ionization zone is maximum,

  - but this luminosity is *delayed in reaching the surface* because of opacity in the hydrogen ionization zone
  - thus the time of maximum surface luminosity occurs *after* the time of minimum radius.

The ε-Mechanism and Stability of Massive Stars

Before the κ-mechanism was proposed it was suggested that stellar pulsations could be driven by variations in the thermonuclear energy production caused by radial oscillations.

- This was called the ε-*mechanism.*
- Just as oscillations can be driven by the κ-mechanism if opacity increases upon contraction, the ε-mechanism can enhance oscillations if *energy production increases upon contraction* (a condition that is usually satisfied).
- Although oscillations can alter the thermonuclear energy production by causing density and temperature variations, this is of importance only in the more central regions of the star where energy production is taking place.
- The problem then is that in the central regions the amplitudes of fundamental modes and overtones are small, making it difficult for changes there to drive oscillations strongly enough to sustain them.
- Thus the ε-mechanism is not likely to be significant for most variable stars.
- However, it is thought that it may be important for the stability of very massive stars (of order $100\ M_\odot$), where oscillations coupled to variations in energy production deep in the star may generate pulsations causing the star to shed surface layers.

15.4 Non-Radial Pulsation

For the variable stars in Table 15.1 that are labeled NR, the mode of pulsation is not spherically symmetric. The corresponding oscillations are called *non-radial modes.*

- Stars exhibiting non-radial pulsation include the δ Scuti stars, β Cephei stars, and ZZ Ceti stars.
- In addition, our own Sun is not presently classified as a variable star (it presumably will become variable after it leave the main sequence and passes through the instability strip in the HR diagram),
- but it undergoes weak non-radial pulsations that are the target of the helioseismology observations described earlier.
- Such non-radial pulsations are somewhat beyond the scope of our present discussion.

Chapter 16

White Dwarfs and Neutron Stars

Red giants will eventually consume all their accessible nuclear fuel.

- After ejection of the envelope, the cores of these stars shrink to the very hot, very dense objects that we call *white dwarfs.*
- An even more dense object termed a *neutron star* can be left behind after the evolution of more massive stars terminates in a core-collapse supernova explosion.

Technically, white dwarfs and neutron stars are stellar corpses, not stars, but it is common to refer to them loosely as stars.

16.1 Sirius B

The bright star Sirius, in Canis Major, is actually a double star.

- The brighter component is labeled Sirius A and the fainter companion star is known as Sirius B.
- Sirius B is an example of a white dwarf.
- Because of its proximity to Earth, Sirius B is not particularly dim (visual magnitude $m_{\rm V} = 8.5$), but it is difficult to observe because it is so close to Sirius A.
- Sirius B is clearly not a normal star; its spectrum and luminosity indicate that it is hot (about 25,000 K surface temperature) but very small.
- This spectrum contains *pressure-broadened hydrogen lines*, implying a surface environment with *much higher density* than that of a normal star.

- Assuming the spectrum of Sirius B to be blackbody and using the well-established distance to Sirius,
- we conclude from its luminosity that Sirius B has a radius of only about 5800 km.
- But Sirius is a visual binary with a very well studied orbit.
- Therefore, we may use Kepler's laws to infer that the mass of Sirius B is about 1.03 $M_\odot$.
- We conclude that a white dwarf like Sirius B packs the mass of a star in an object the size of the Earth.

Sirius B is the nearest and brightest white dwarf and we shall often use it as illustration.

- However, it is in some respects not so representative because
- its mass of about 1.03 $M_\odot$ is much larger than the average mass of about $0.58 M_\odot$ observed for white dwarfs.

16.2 Properties of White Dwarfs

The preceding discussion allows us to make some immediate estimates that will shed light on the nature of white dwarfs even before we carry out any detailed analysis.

16.2.1 Density and Gravity

- Since white dwarfs contain roughly the mass of the Sun in a sphere the size of the Earth, we expect that white dwarfs have densities in the vicinity of $10^6\,\mathrm{g\,cm^{-3}}$.
- For Sirius B the average density calculated from the observed mass and radius is about $2.5\times10^6\,\mathrm{g\,cm^{-3}}$.
- The gravitational acceleration and the escape velocity at the surface for Sirius B are

  $$g = \frac{Gm}{R^2} \simeq 3.7\times10^8\,\mathrm{cm\,s^{-2}} \qquad \frac{v_{\mathrm{esc}}}{c} = \sqrt{\frac{2Gm}{Rc^2}} \simeq 0.02,$$

  respectively, indicating that

  - the gravitational acceleration is almost 400,000 times larger than at the Earth's surface, but
  - general relativity effects, while not completely negligible, are sufficiently small to be ignored in initial approximation.

16.2.2 Equation of State

- We conclude from the preceding that hydrostatic equilibrium under Newtonian gravitation is adequate as a first approximation for the structure of white dwarfs.
- What about the microphysics of the gas?
  - Can we apply a Maxwell–Boltzmann description, or will the quantum statistical properties of the gas play a crucial role?
  - Will electron velocities be describable classically or will velocities become relativistic?

Let's assume nonrelativistic velocities and that electrons are responsible for the internal pressure of the white dwarf.

- For simplicity we assume that the white dwarf is composed of a single kind of nucleus having atomic number Z, neutron number N, and atomic mass number $A = Z + N$.
- Then the average electron velocity is $\bar{v}_{\mathrm{e}} = \bar{p}/m_{\mathrm{e}}$ where $\bar{p}$ is the average momentum and m_{e} is the electron mass.
- By the *uncertainty principle*, the average momentum is

  $$\bar{p} \simeq \Delta p \simeq \hbar/\Delta x \simeq \hbar n_{\mathrm{e}}^{1/3},$$

  where n_{e} is the electron number density.
- We may expect the gas to be completely ionized and the corresponding electron number density is

  $$n_{\mathrm{e}} = \left(\frac{\text{number } e^-}{\text{nucleon}}\right)\left(\frac{\text{number nucleons}}{\text{unit volume}}\right) = \left(\frac{Z}{A}\right)\left(\frac{\rho}{m_{\mathrm{H}}}\right).$$

- Therefore, the average electron velocity is

  $$\frac{\bar{v}_{\mathrm{e}}}{c} = \frac{\bar{p}}{m_{\mathrm{e}}c} = \frac{\hbar n_{\mathrm{e}}^{1/3}}{m_{\mathrm{e}}c} = \frac{\hbar}{m_{\mathrm{e}}c}\left(\frac{Z\rho}{Am_{\mathrm{H}}}\right)^{1/3} \simeq 0.25,$$

  where we assume that $A = 2Z$, as would be true for ^{12}C, ^{16}O, or ^{4}He (primary constituents of most white dwarfs).

We conclude that *electron velocities will become relativistic* for higher-density white dwarfs.

- The *average spacing* between electrons in the gas is

  $$d \simeq n_{\rm e}^{-1/3} \simeq 1.5 \times 10^{-10}\,{\rm cm},$$

  using $Z/A = 0.5$ and the average density of Sirius B.

- The average *deBroglie wavelength* of the electrons is

  $$\bar{\lambda}_{\rm e} = \frac{h}{\bar{p}} = \frac{h}{m_{\rm e}\bar{v}_{\rm e}} \simeq 9.6 \times 10^{-10}\,{\rm cm}.$$

- Since $d < \bar{\lambda}_{\rm e}$, the *electron gas will be degenerate*, provided that the temperature is not too high.

- For a degenerate fermion gas the *fermi energy* is

  $$E_{\rm f} = \sqrt{k_{\rm f}^2 + m^2} \qquad (\hbar = c = 1).$$

  The gas remains degenerate as long as $E_{\rm f} \gg kT$.

- From the preceding equation $E_{\rm f} \geq m_{\rm e}c^2 = 0.511\,{\rm MeV}$, and

  $$T = E/k > 0.511\,{\rm MeV}/k \simeq 6 \times 10^9\,{\rm K}$$

  is required to *break the degeneracy*.

- Simulations indicate that interior white dwarf temperatures are typically 10^6–10^7 K, so we conclude that *white dwarfs contain cold, degenerate gases of electrons.*

- Thus they approximated by *polytropic equations of state*,

  $$P = K\rho^{\gamma},$$

  1. $\gamma = \frac{5}{3}$ for nonrelativistic degenerate electrons
  2. $\gamma = \frac{4}{3}$ for ultrarelativistic degenerate electrons.

- While we expect the electrons to be degenerate and to become relativistic at higher densities, the ions are much more massive than the electrons.

- The ions are neither relativistic nor degenerate, and are well described by an ideal gas equation of state.

- Ions move slowly so they contribute little pressure.

- However, calculations indicate that most of the heat energy stored in the white dwarf is associated with motion of the ions.

- Finally, photons constitute a relativistic gas approximated by a Stefan–Boltzmann equation of state,

  $$P = \tfrac{1}{3}aT^4,$$

  where T is the temperature.

A white dwarf is a hot, dense object for which

- *mechanical properties* (like pressure, generated mostly by the degenerate electrons)

- are *decoupled* from the *thermal properties* (which are associated primarily with the ions at normal temperatures).

16.2.3 Ingredients of a White Dwarf Description

An initial description of a white dwarf requires a theory where

1. Stable configurations correspond to hydrostatic equilibrium under Newtonian gravitation.
2. Ions carry most of the mass and store most of the thermal energy, but electrons provide most of the pressure.
3. The electron equation of state is that of a *cold degenerate gas*, approximated as $P = K\rho^{\gamma}$, with $\gamma = \frac{5}{3}$ for nonrelativistic and $\gamma = \frac{4}{3}$ for relativistic electrons, respectively.
4. Ions constitute a nonrelativistic ideal gas.
5. Photons obey a Stefan–Boltzmann equation of state.
6. Because the degenerate electron gas is primarily responsible for the pressure but its equation of state does not depend on temperature, the *thermal and mechanical properties of the white dwarf are decoupled*.
7. As density increases the velocity of the electrons increases and special relativity becomes important, corresponding to a transition

   $$P \simeq K\rho^{5/3} \longrightarrow P \simeq K'\rho^{4/3}.$$

   in the electron equation of state.

Let us now turn to a theoretical description embodying these basic ideas in a relatively simple formulation.

16.3 Polytropic Models of White Dwarfs

We expect that white dwarfs are approximately described by systems in hydrostatic equilibrium having degenerate electron equations of state.

- Thus, we may expect that solutions of the Lane–Emden equation with polytropic index $n = \frac{3}{2}$, corresponding to $\gamma = \frac{5}{3}$, are relevant for the structure of low-mass white dwarfs where electron velocities are nonrelativistic.
- Likewise, we may expect that in more massive white dwarfs the electrons become relativistic and the corresponding structure is related to a Lane–Emden solution with polytropic index $n = 3$, corresponding to $\gamma = \frac{4}{3}$.
- Between these extremes the electron equation of state must generally be described in numerical terms permitting an arbitrary level of degeneracy and degree of relativity.

16.3.1 Low-Mass White Dwarfs

Let us first consider a low-mass white dwarf.

- Assuming a $\gamma = \frac{5}{3}$ polytropic equation of state ($n = \frac{3}{2}$), the relationship of the mass M and radius R is given by the Lane–Emden result (see Ch 8)

$$M = 4\pi R^{(3-n)/(1-n)} \left(\frac{(n+1)K}{4\pi G}\right)^{n/(n-1)} \xi_1^{(3-n)/(n-1)} \xi_1^2 |\theta'(\xi_1)|.$$

  which implies that

$$MR^3 = \text{constant},$$

  since

$$R^{(3-n)/(1-n)} = R^{(3-3/2)/(1-3/2)} = R^{(3/2)/(-1/2)} = R^{-3}.$$

- Thus the *product of the mass and the volume of a low-mass white dwarf is constant*.

> We obtain the surprising result that, contrary to the behavior of normal stars, increasing the mass of a low-mass white dwarf causes its radius to *shrink*.

- This behavior is a direct consequence of a *degenerate electron equation of state*.

16.3.2 The Chandrasekhar Limit

If we continue to add mass to a white dwarf, the electrons eventually will become relativistic ($\gamma = \frac{4}{3}$ or $n = 3$), and

$$M = 4\pi R^{(3-n)/(1-n)} \left(\frac{(n+1)K}{4\pi G}\right)^{n/(n-1)} \xi_1^{(3-n)/(n-1)} \xi_1^2 |\,\theta'(\xi_1)|.$$

then implies that

$$M = \text{constant} \times R^0 = \text{constant}$$

This even more surprising result defines the *Chandrasekhar limiting mass*, which implies that there is an *upper limit for the mass of a white dwarf.*

Inserting the constants, we find for a high-mass white dwarf

$$R = 3.347 \times 10^4 \left(\frac{\rho_c}{10^6\,\text{g}\,\text{cm}^{-3}}\right)^{-1/3} \left(\frac{\mu_e}{2}\right)^{-2/3} \text{km},$$

and for the *Chandrasekhar mass*,

$$M_0 = 1.457 \left(\frac{2}{\mu_e}\right)^2 M_\odot \simeq 1.4 M_\odot,$$

where the last estimate follows because generally $2/\mu_e \sim 1$.

Thus the Chandrasekhar limiting mass is slightly composition dependent but implies an upper mass limit for a white dwarf of approximately 1.4 $M_\odot$.

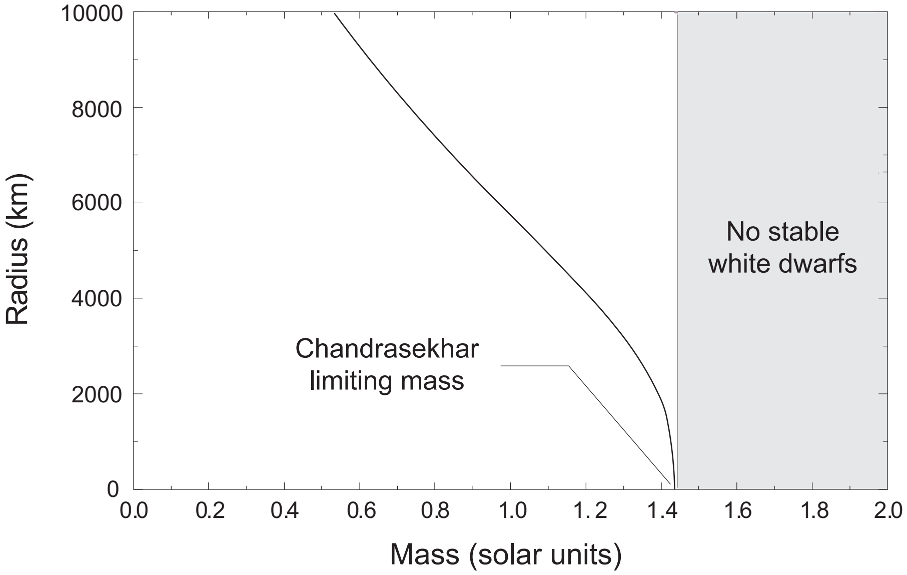
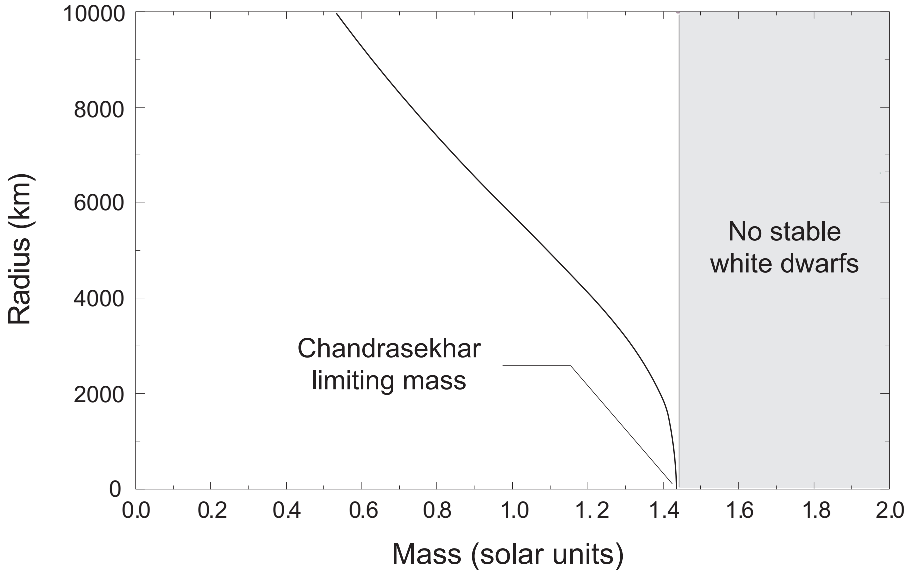


Figure 16.1: Dependence of radius on mass for a white dwarf. The Chandrasekhar limit of 1.44 solar masses is indicated. This calculation assumes an electron fraction of $Y_e = 0.5$. (The electron fraction Y_e is the ratio of the number of electrons to the total number of nucleons. For symmetric matter $Z = N$, so for fully-ionized symmetric matter, $Y_e = \frac{1}{2}$.) Thus, for electrons this equation of state approximates a $\gamma = \frac{5}{3}$ polytrope at low mass and a $\gamma = \frac{4}{3}$ polytrope at high mass, with a smooth transition in between. Ions of the white dwarf are assumed to obey an ideal gas equation of state and the photons are described by a Stefan–Boltzmann photon gas equation of state.

- In Fig. 16.1 the radius versus mass for white dwarfs in hydrostatic equilibrium is shown for a numerical simulation.
- This calculation uses a numerical equation of state that accounts fully for arbitrary degrees of electron degeneracy and arbitrary relativity for electrons.
- Thus, for electrons this equation of state approximates a $\gamma = \frac{5}{3}$ polytrope at low mass and a $\gamma = \frac{4}{3}$ polytrope at high mass, with a smooth transition in between.



- The ions of the white dwarf are assumed to obey an ideal gas equation of state and the photons are described by a Stefan–Boltzmann photon gas equation of state.

- The above figure shows the behavior implied by the preceding equations.

  1. For lower masses the radius of the white dwarf decreases steadily with increase in mass, in accord with $MR^3 = \text{constant}$, but
  2. At high masses the curve approaches a vertical asymptote given by $M = M_0$, with the calculation becoming numerically unstable near the limiting mass.

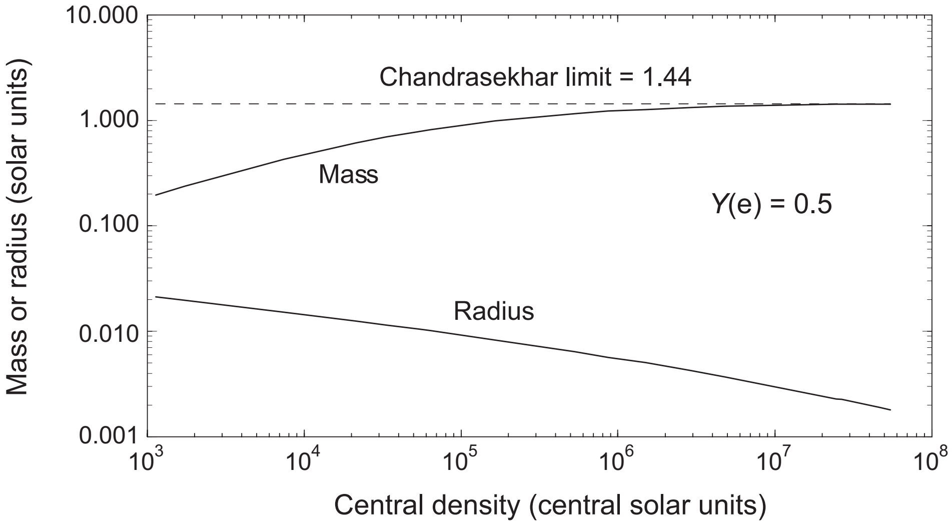


Figure 16.2: The variation of mass and radius for white dwarfs as a function of the central density in units of central solar densities.

- In Fig. 16.2 the variation of the mass and radius of white dwarfs as a function of the central density in central solar units is plotted for calculations similar to those described in Fig. 16.1.
- Note the steady trend to zero radius as the white dwarf approaches the limiting mass asymptotically.

16.3.3 Heuristic Derivation of the Chandrasekhar Limit

The Chandrasekhar limiting mass was obtained above as a consequence of the Lane–Emden equations, which embody

1. Polytropic equations of state and
2. Hydrostatic equilibrium.

It will prove useful in understanding the limiting mass for white dwarfs to obtain the Chandrasekhar result in a somewhat more intuitive way.

- Assume a fully-ionized sphere of symmetric *(equal numbers of protons and neutrons)* matter containing N electrons.
- The mass of the sphere is then $M \simeq 2m_{\rm p}N$,
- the average spacing between electrons is $d \sim R/N^{1/3}$, and
- the average momentum of the electrons is (uncertainty principle)
$$p_{\rm f} \sim \frac{\hbar}{d} \sim \frac{\hbar M^{1/3}}{R m_{\rm p}^{1/3}}.$$
- Estimate the total energy of the degenerate electrons and balance that against the gravitational energy of the protons.
- This gives in the nonrelativistic and relativistic limits,
$$E = a\frac{M^{5/3}}{R^2} - b\frac{M^2}{R} \qquad \text{(nonrelativistic)}$$
$$E = c\frac{M^{4/3}}{R} - d\frac{M^2}{R} \qquad \text{(relativistic)}$$
where a, b, c, and d are positive constants.
- Notice that the two terms in the nonrelativistic case have different dependence on R.
- Thus, by setting $\partial E/\partial R = 0$, we find an equilibrium configuration in the nonrelativistic case that generally satisfies
$$MR^3 = \text{constant}.$$

- On the other hand, in the relativistic case

$$E = c\frac{M^{4/3}}{R} - d\frac{M^2}{R} \qquad \text{(relativistic)}$$

  the two terms have the *same* dependence on R.

- Thus, trying to solve $\partial E/\partial R = 0$ for R corresponding to a stable configuration leads to an *indeterminate result* (the resulting equation does not depend on R).

- Note that both terms in this equation vary as R^{-1}, but the first term depends on $M^{4/3}$ while the second varies as M^2.

- The second term has a net negative sign and a stronger dependence on M than the first term, so *the total energy becomes negative* if the mass is made large enough.

- But the total energy scales as R^{-1}, so once the total energy becomes negative *the energy can be minimized by shrinking to zero radius:*

> *For a relativistic degenerate gas, exceeding a limiting mass leads to gravitational collapse.*

- We may estimate this critical mass by equating the two terms in $E = cM^{4/3}/R - dM^2/R$, yielding

$$M_0 = \left(\frac{\hbar c}{G m_{\mathrm{p}}^{4/3}}\right)^{3/2} \simeq 1\, M_\odot,$$

  which is correct to order of magnitude.

16.3.4 The Adiabatic Index γ and Gravitational Stability

The preceding results are another variation of the theme introduced previously in conjunction with the collapse of protostars.

- There we found that an adiabatic index of $\gamma \le \frac{4}{3}$ implies an instability against expansion or contraction.
- From the polytropic equation of state $P = K\rho^\gamma$,

  $$\frac{dP}{d\rho} = K\gamma\rho^{\gamma-1} \;\rightarrow\; \frac{\rho}{P}\frac{dP}{d\rho} = \gamma\frac{\rho}{P}K\rho^{\gamma-1} = \gamma\frac{K\rho^\gamma}{P} = \gamma$$

  suggesting that we define γ for any equation of state $P(\rho)$ by

  $$\gamma \equiv \frac{\rho}{P}\frac{dP}{d\rho} = \frac{d\ln P}{d\ln\rho}.$$

- Taking this logarithmic derivative as the definition of an effective adiabatic index γ_{eff}, we may expect that in any simulation of hydrostatic equilibrium,

  $$\gamma_{\text{eff}} \equiv \frac{\rho}{P}\frac{dP}{d\rho} \simeq \frac{4}{3}$$

  heralds the onset of a radial scaling instability.

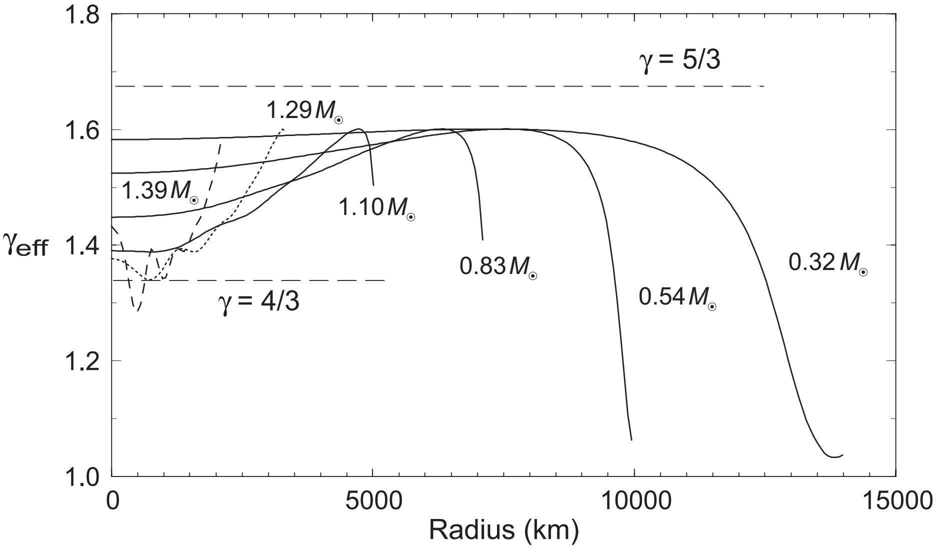
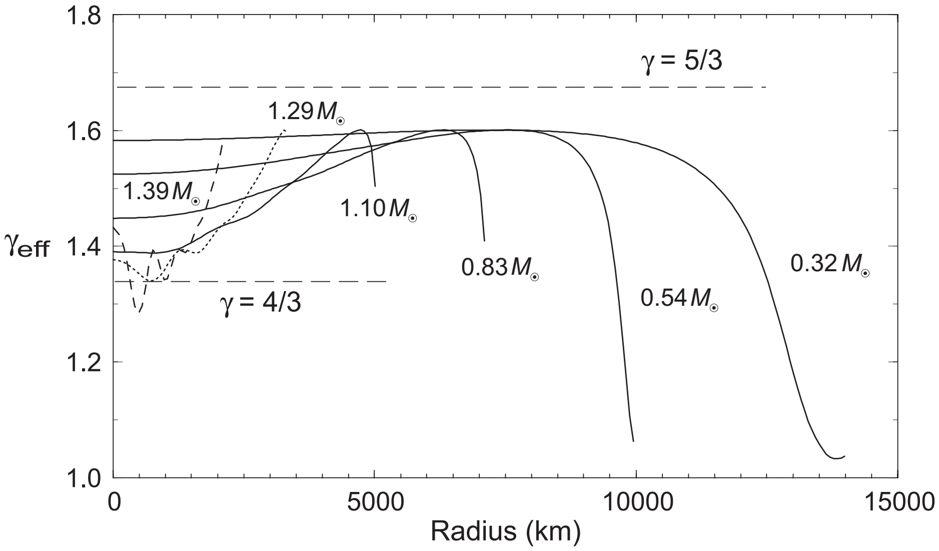


Figure 16.3: Values of the parameter $\gamma \equiv d\ln P/d\ln\rho$ at constant temperature for white dwarfs of various masses (solar units). The values of γ corresponding to nonrelativistic ($\gamma = 5/3$) and relativistic ($\gamma = 4/3$) polytropes are indicated. The calculation becomes unstable as the mass approaches the Chandrasekhar limiting mass which is 1.44 solar masses for this calculation (for which $Y_e = 0.5$). The central temperature is assumed to be 5×10^6 K in all calculations.

- In Fig. 16.3 the value of γ_{eff} as a function of radius is calculated numerically using

$$\gamma = \frac{\rho}{P}\frac{dP}{d\rho} = \frac{d\ln P}{d\ln\rho}.$$

for white dwarf solutions that have been obtained with an equation of state allowing arbitrary electron degeneracy and relativity.



- For low-mass white dwarfs the effective value of γ is near the nonrelativistic expectation of $\frac{5}{3}$ for the entire interior.

- However, as the mass of the white dwarf is increased, the effective value of γ in the deep interior begins to drop.

- As the mass approaches the Chandrasekhar limit, $\gamma_{\rm eff} \to \frac{4}{3}$ and the numerical solution becomes very unstable.

- These numerical fluctuations reflect the incipient gravitational instability that in this case occurs at 1.44 solar masses.

The roles of relativity and quantum mechanics are central to the preceding results.

- Nonrelativistic degenerate matter has $\gamma \sim \frac{5}{3}$, which is gravitationally stable.
- But quantum mechanics (the uncertainty principle) requires the electrons to move faster as the density increases, implying that the velocities eventually become relativistic as the white dwarf mass increases.
- Relativistic degenerate matter has $\gamma \sim \frac{4}{3}$, which inherently is gravitationally unstable.
- Because the speed of the electrons is limited by the speed of light, there is a mass beyond which even the degeneracy pressure cannot prevent gravitational collapse of the system.

This critical point is the Chandrasekhar limiting mass.

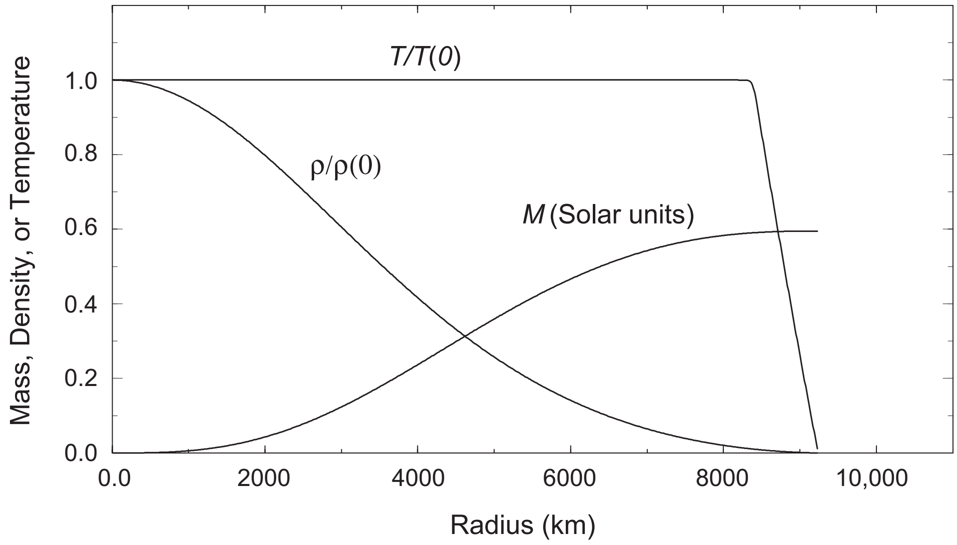


Figure 16.4: Behavior of density, enclosed mass, and temperature for a white dwarf. In this calculation the white dwarf has a central density of $2.9 \times 10^6\ \mathrm{g\,cm^{-3}}$, a central temperature of 5.0×10^6 K, a total mass of 0.595 solar masses, and a radius of 9234 km.

16.4 Internal Structure of White Dwarfs

- A numerical calculation of the internal structure of a white dwarf is illustrated in Fig. 16.4, which plots the density, enclosed mass, and temperature as a function of radius.
- The calculation corresponds to hydrostatic equilibrium with a realistic electron equation of state in which the electrons have arbitrary degeneracy and degree of relativity.
- The ions are assumed to obey an ideal gas equation of state and radiation to obey a Stefan–Boltzmann equation of state.

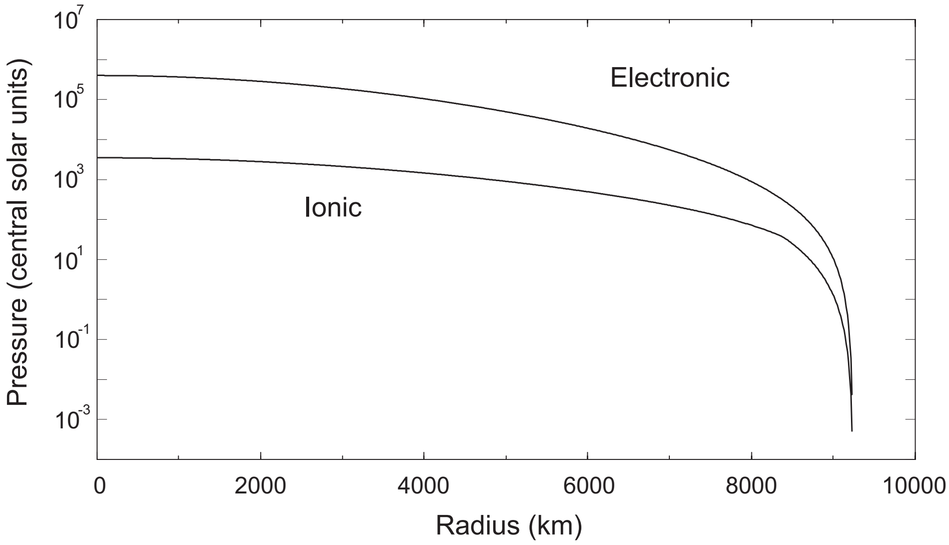
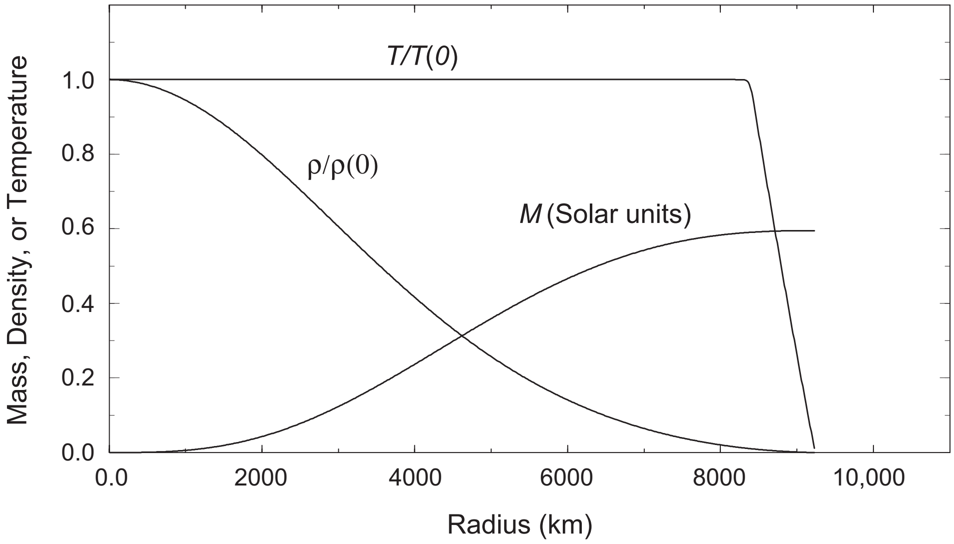


Figure 16.5: Relative contributions of the electronic pressure and ionic pressure for the calculation described in Fig. 16.4. The contribution to the pressure from radiation under these conditions is completely negligible relative to the electronic and ionic contributions. The electronic contribution is very nearly that expected for a fully degenerate gas.

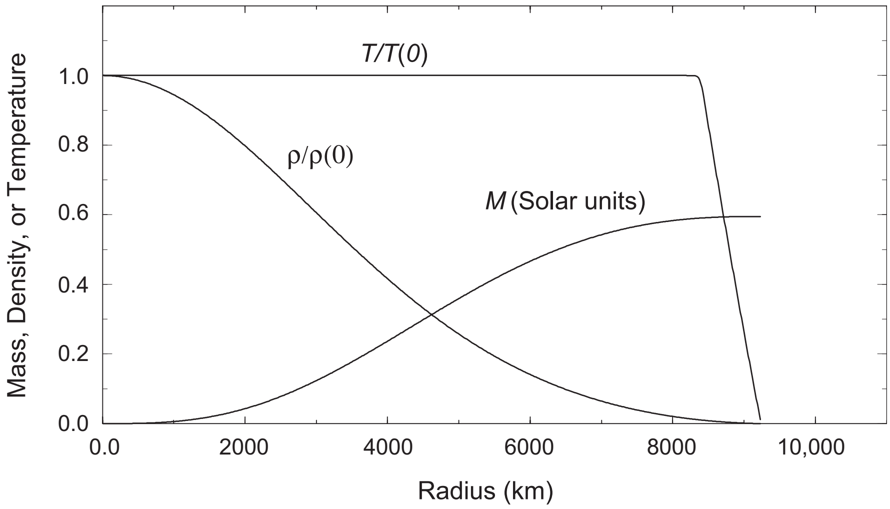
- Figure 16.5 illustrates the relative contribution of
  - electrons and
  - ions

  to the pressure in the preceding calculation.
- It provides strong justification for our earlier assumption that the pressure in white dwarfs is dominated by the contribution from degenerate electrons.



The internal temperature variation in the calculation shown above is determined as follows.

- Degenerate matter is a good conductor of thermal energy.
- Thus the interior of a white dwarf cannot support a substantial temperature gradient and
- we assume all but a thin surface layer to be isothermal and strongly heat conducting.
- Near the surface the density drops to zero and the nearly ideal gas expected there is a very good insulator.
- This suggests that a good model of how white dwarfs cool is one of a conducting sphere with no temperature gradient surrounded by a thin layer of normal gas with a gradient set by its transport properties (that is, by its opacity).
- This model is analogous mathematically to cooling of a hot metal ball surrounded by a thin insulating jacket, since degenerate gases have many of the properties of metals.



In the "metal ball plus insulating blanket" model for the above figure,

- The interior is assumed fully conductive,
- The surface is assumed insulating with a radiative opacity given by the Kramers bound–free opacity, and
- the transition between the two is governed by the degeneracy parameter

$$\alpha = \frac{\mu - m_e c^2}{kT},$$

where μ is the chemical potential for the electrons.

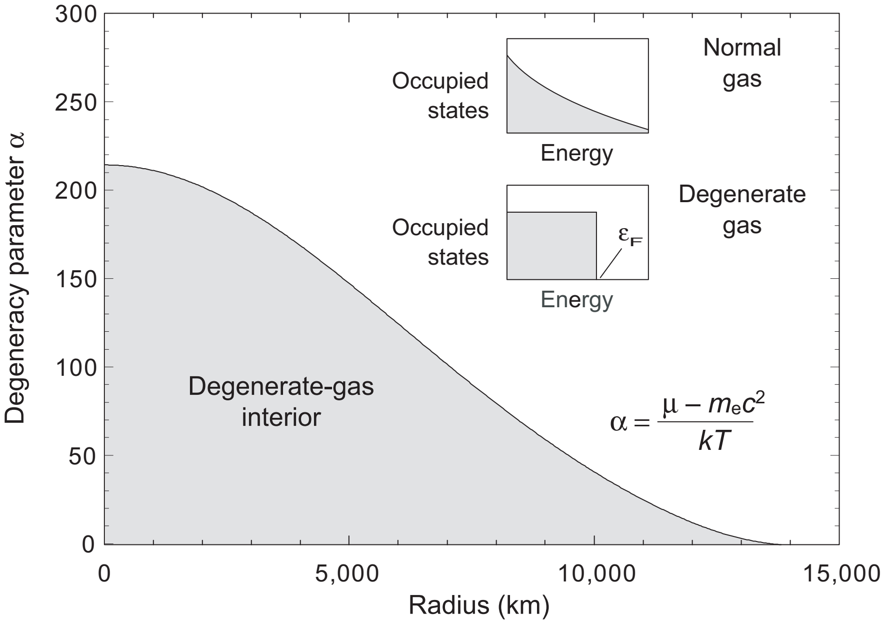


Figure 16.6: The degeneracy parameter $\alpha \equiv (\mu - m_e c^2)/kT$ versus radius for a white dwarf simulation, where μ is the electron chemical potential and m_e the electron mass. A similar equation of state as for Fig. 16.1 was used in the calculation. We see that the electron gas is highly degenerate except very near the surface of the star. Shown inset are occupation profiles for a normal gas and a degenerate electron gas, with ε_F the Fermi energy.

- The variation of the degeneracy parameter

$$\alpha = \frac{\mu - m_e c^2}{kT},$$

  with radius is illustrated in Fig. 16.6.
- In the interior α is large, indicating high degeneracy.
- But very near the surface α falls to zero, implying that in a thin surface layer the electrons obey approximately an ideal gas law.

16.5 Cooling of White Dwarfs

Although white dwarfs have no internal heat source, they can remain luminous for long periods of time as the heat left over from their glory days slowly leaks away.

- The cooling curve for a white dwarf should then reflect both the internal structure and the age of the star.
- As we have seen, white dwarfs are well described by a spherical ball of electron degenerate matter surrounded by a very thin surface layer that obeys an ideal gas equation of state.

This can serve as a simple but quantitative model for cooling rates in white dwarfs.

- By determining observationally the surface temperature of white dwarfs in a stellar population and relating these to theoretical cooling curves,
- it is possible to estimate the age of the white dwarfs and hence infer the age of the stellar population.

Such methods are used extensively to determine the age of stellar populations in our galaxy.

- White dwarfs may cool by neutrino emission from hot, dense central regions in addition to cooling by photon emission from the surface.
- This is in fact thought to be the dominant source of cooling for young, hot white dwarfs and occurs primarily through emission of plasma neutrinos from the deep interior.

Neutrino Cooling of White Dwarfs:

White dwarfs can cool by emission of neutrinos from the interior as well as through photons emitted from the surface.

- The dominant source of neutrino cooling is expected to be plasma neutrinos emitted from the central region, for white dwarfs with surface temperatures greater than about $25,000$ K.
- It has been proposed that neutrino emission might be observed *indirectly* by studying the effect of neutrino cooling on pulsations of young, hot, variable white dwarfs.
- The DBV white dwarfs (white dwarfs with a helium atmosphere that are pulsating variables) have effective surface temperatures around 25,000 K,
- so they are thought to cool largely through emission of plasma neutrinos.
- Simulations indicate that the rate of change in the observed pulsation period versus time is affected significantly by neutrino emission.

This suggests that changes observed in the pulsation period of a suitable DBV white dwarf could be used to infer the rate of neutrino cooling.

16.6 Crystallization of White Dwarfs

In the early 1960s it was predicted that as the plasma in a white dwarf cools

- it may become energetically favorable for the ions to form a *body-centered cubic (BCC) crystalline lattice* to minimize the Coulomb repulsion.
- This is expected to occur through a first-order *liquid to solid phase transition*.
- The corresponding *latent heat of crystallization* provides a new energy source.
- This supplements the thermal energy stored in the ions and influences the subsequent thermal evolution of the white dwarf.
- Whether this transition occurs, and its detailed properties if it does, constitutes one of the largest uncertainties in calculating white dwarf cooling.
- This in turn has implications for the use of white dwarf cooling curves to determine the age of stellar populations.
- It is possible to study the internal structure of some stars through *asteroseismology,* by extending the helioseismology concepts used to study the Sun.

These methods provide a way to test the hypothetical crystallization of cooling white dwarfs.

- For typical white dwarfs theory suggests that crystallization in the core begins when the surface temperature decreases to 6000–8000 K.
- However, in more massive white dwarfs crystallization is expected to set in at a higher surface temperature.
- Thus asteroseismology on massive white dwarfs is a promising source of evidence for crystallization.
- Asteroseismology of the pulsating DAV white dwarf BPM 37093 has been used to infer its internal structure.
- This star represents a particularly favorable case because its mass of $1.1 M_\odot$ is the largest known for a DAV white dwarf.
- The oscillations of this and other pulsating white dwarfs correspond to non-radial gravity waves (g-modes), which represent oscillations with a restoring force provided by gravity.
- If the core of a white dwarf becomes solid because of the crystalline phase transition, the difference in density at the solid–liquid core boundary is very small.

Thus the mechanical properties of the white dwarf are not altered significantly and the effect on evolution of the white dwarf is expected to be minimal.

- However, formation of a crystalline core may have a *significant effect on the star's pulsations* because
  - the additional shear in the solid relative to the liquid causes a mismatch between interior and exterior waves at the core boundary, and
  - the exterior waves are almost completely reflected by the boundary.
- Hence, the nonradial g-modes
  - *cannot penetrate* the solid–liquid interface,
  - the white dwarf's observable pulsations become linked to g-modes confined to the non-crystalline liquid region outside the core, and
  - the size of the crystalline core exerts a potentially-observable effect on the pulsations of the star.

From analysis of the observed pulsation frequencies

- it was concluded that BPM 37093 has a core of crystallized carbon and oxygen containing about 90% of the white dwarf's mass.
- A different analysis of BPM 38093 observational data concluded that the crystalline mass most likely lies between 32% and 82%.
- In either case there is credible evidence that *a substantial fraction of the white dwarf has entered the crystal phase* predicted by theory.
- Most white dwarfs are rich in carbon so crystallized white dwarfs have been referred to whimsically as *"diamonds in the sky"*.

Accordingly, BPM 37093 has been nicknamed *"Lucy"* by some, in reference to the famous Beatles song *Lucy in the Sky with Diamonds.*

16.7 Beyond White Dwarf Masses

The preceding discussion of limiting masses for white dwarfs assumes all pressure to derive from electrons.

- However, if the Chandrasekhar mass is exceeded and the system collapses, eventually a density will be reached where the nucleons (also fermions) will begin to produce a strong degeneracy pressure.
- Whether this nucleon degeneracy pressure can halt the collapse depends on the mass.
- Calculations indicate that for a mass less than about 2–3 solar masses (depending weakly on details such as the equation of state), the collapse converts essentially all protons into neutrons through the weak interactions, producing a neutron star.
- The degeneracy pressure of the neutrons halts the collapse at neutron-star densities and radii approximately 500 times smaller than for white dwarfs.
- Calculations, and general considerations for strong gravity, indicate that for masses greater than this even the neutron degeneracy pressure cannot overcome gravity and the system collapses to a black hole.
- These considerations also indicate that white dwarfs and neutron stars are the only possible stable configurations lying between normal stars and black holes.

Therefore, let us now consider neutron stars.

16.8 Basic Properties of Neutron Stars

- Neutron stars were predicted in 1933 by Baade and Zwicky as a possible end result of what we would now call a core-collapse supernova.
- Oppenheimer and Volkov worked out equations describing their general structure and properties in 1939. *(Requires general relativity)*
- However, they were not taken very seriously until the discovery of radio pulsars in the 1960s pointed to rapidly rotating neutron stars as their most likely explanation.
- Now thousands have been observed.

- Most neutron stars have been discovered as radio pulsars but the vast majority of the energy emitted by neutron stars is in very high-energy photons (X-rays and γ-rays), rather than radio waves.
- Typically only about 10^{-5} of their radiated energy is in the radio-frequency spectrum.
- Most neutron stars have masses of 1–2 $M_\odot$ and diameters of 10–20 km. Very loosely, a neutron star packs the mass of a normal star like the Sun into a volume of order 10 km in radius.
- From the density of a little over 1 $\mathrm{g\,cm^{-3}}$ and radius of about 7×10^5 km for the Sun, we may estimate immediately an average density of order $10^{14}\,\mathrm{g\,cm^{-3}}$ for neutron stars (it can actually be about an order of magnitude larger than that).
- Thus, they have enormous densities that are similar to those encountered in the nucleus of the atom.
- In fact, in certain ways (but not all), neutron stars are similar to giant atomic nuclei the size of a city.
- Their enormous densities imply strong gravitational fields and the possibility of significant general relativistic deviations from Newtonian gravity.

Electron Capture and Neutronization

The formation of a neutron star results from a process called *electron capture* (a form of beta decay), which can follow the core collapse of a massive star late in its life to produce a supernova (see further Ch. 15).

- The process is also called *neutronization,* because its effect is to destroy protons and electrons and create neutrons. The basic reaction is

$$e^- + p^+ \to n^0 + \nu_\mathrm{e}.$$

- It is slow under normal conditions (because it is mediated by the weak interaction), but very fast in the high density and temperature environment produced by core collapse in a massive star.
- In the supernova explosion the enormous amount of energy released gravitationally in the collapse of the core blows off the outer layers of the star and leaves behind an extremely dense, hot remnant.
- As the neutronization reaction proceeds, the neutrinos escape carrying off energy and leave behind the neutrons.
- Because neutrons carry no charge, there is no electrical repulsion as in normal matter and the core can collapse to very high density once it has become mostly neutrons.
- The structure of actual neutron stars is more complex than this, and they are not composed entirely of neutrons, but this simple picture captures the basic idea.

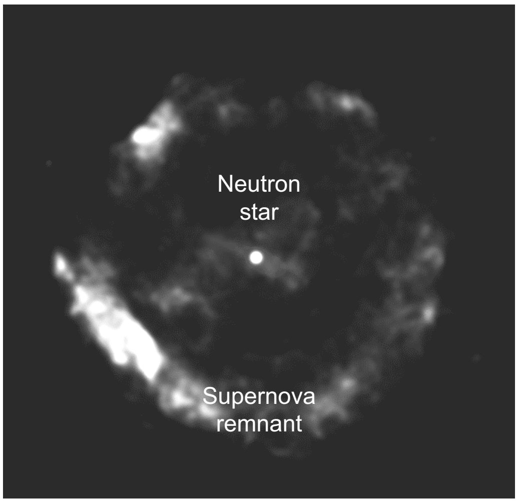


Figure 16.7: Chandra X-ray observatory image of a neutron star in the center of an expanding supernova remnant. This neutron star is also a pulsar.

Because neutron stars are tiny it might be expected that they would be very difficult to detect.

- In fact, neutron stars have luminosities that are comparable to that of stars like the Sun because they have very high surface temperatures (of order 10^6 K).
- Because of the high temperature, the light emitted peaks in the extreme-UV and soft X-ray portion of the spectrum,
- so neutron stars are readily visible to X-ray observatories.
- An X-ray image of a neutron star at the center of an expanding supernova remnant is shown in Fig. 16.7.

It is believed that the neutron star and the expanding remnant surrounding it were produced by a supernova seen on Earth in 386 AD by Chinese observers.

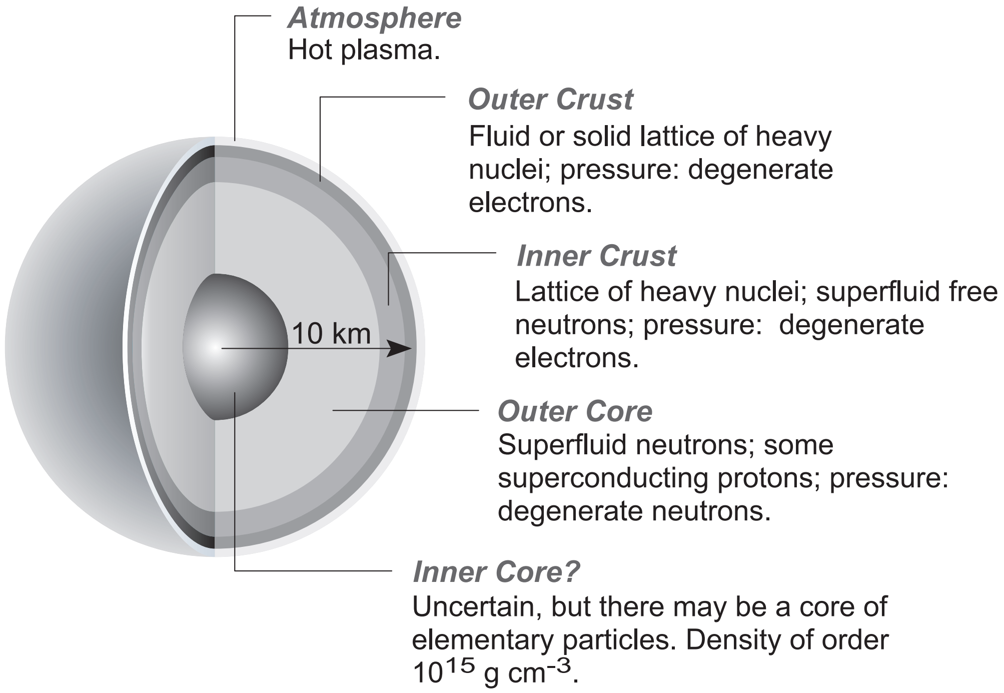
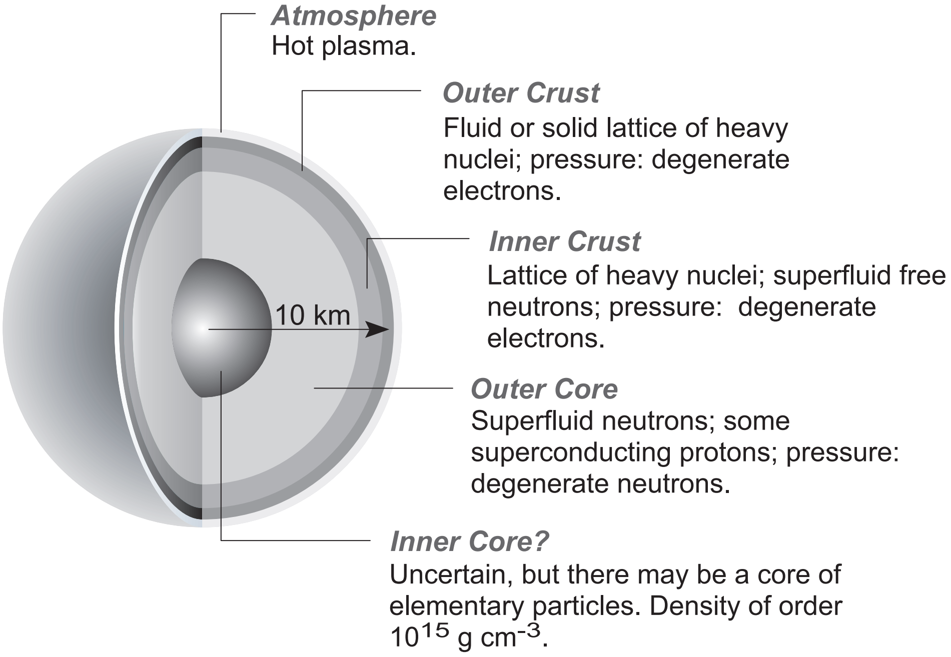

Atmosphere
Hot plasma.
Outer Crust
Fluid or solid lattice of heavy nuclei; pressure: degenerate electrons.
Inner Crust
Lattice of heavy nuclei; superfluid free neutrons; pressure: degenerate electrons.
10 km
Outer Core
Superfluid neutrons; some superconducting protons; pressure: degenerate neutrons.
Inner Core?
Uncertain, but there may be a core of elementary particles. Density of order 10^{15} g cm^{-3}.

Figure 16.8: Internal structure of a typical neutron star.

Internally, we believe that a neutron star can be divided into the following general regions (see Fig. 16.8).

- The atmosphere of hot, ionized gas is ~ 1 cm thick.
- The outer crust is about 200 meters thick and consists of a solid lattice or a dense liquid of heavy nuclei. The dominant pressure in this region is from electron degeneracy. The density is not high enough to favor neutronization.
- The inner crust is from $\frac{1}{2}$ to 1 kilometer thick. The pressure is higher and the lattice of heavy nuclei is permeated by free superfluid neutrons that begin to "drip" out of the nuclei. Pressure is mostly from degenerate electrons.


Atmosphere
Hot plasma.
Outer Crust
Fluid or solid lattice of heavy nuclei; pressure: degenerate electrons.
Inner Crust
Lattice of heavy nuclei; superfluid free neutrons; pressure: degenerate electrons.
10 km
Outer Core
Superfluid neutrons; some superconducting protons; pressure: degenerate neutrons.
Inner Core?
Uncertain, but there may be a core of elementary particles. Density of order 10^{15} g cm^{-3}.

- The outer core consists primarily of superfluid neutrons and the neutrons supply most of the pressure through neutron degeneracy, though there are some free superconducting protons. This region gives the neutron star its name.

- The structure of the inner core is less certain because we are less certain about how matter behaves under the intense pressure at the center (that is, the *equation of state* for matter under these conditions is not well understood).

- It might even consist of a solid core of particles more elementary than nucleons (pions, hyperons, quarks, . . .).

Much of a neutron star consists of closely packed neutrons and has some resemblance to a giant atomic nucleus, but it is important to remember that it is gravity, not the nuclear force, that holds a neutron star together (see the following box). .

Neutron Stars Are Bound by Gravity

In some ways a neutron star is like 20-km diameter atomic nucleus, but there is one important difference:

> A neutron star is bound by *gravity,* and the strength of that binding is such that the density of neutron stars is even greater than that of nuclear matter.

- How can the weakest force (gravity) produce an object more dense than atomic nuclei, which are held together by a diluted form of the strongest force?
- The answer: *range and sign of the forces involved.*
  - Gravity is weak, but long-ranged and attractive.
  - The strong nuclear force is short-ranged, acting only between nucleons that are near neighbors.
  - The normally attractive nuclear force becomes repulsive at very short distances. (A neutron star would *explode* if gravity were removed.)
- This is a kind of Tortoise and Hare fable:
  - Gravity is weak, but relentless and always attractive.
  - Thus, over large enough distances and long enough time, gravity—the plodding Tortoise of forces—always wins.

That is why the material in a neutron star can be compressed to such high density by the most feeble of the known forces.

16.8.1 Cooling of Neutron Stars

Neutron stars form from the innermost material left behind in a core collapse supernova.

- The *protoneutron star* formed in the supernova is initially very hot and bloated.
- It typically has $T \sim 10^{11}$ K and a radius some 30% larger than the final neutron star that it will become),
- It is still being powered by accretion from the part of the envelope that did not escape the star in the explosion.

As the accretion tapers off the nascent neutron star cools rapidly by neutrino emission.

- In high-energy astrophysics temperatures are often quoted in energy units,
- with the corresponding temperature in kelvin given by $T = E/k$, where k is Boltzmann's constant.

A characteristic temperature for a protoneutron star is $\sim 50\,\text{MeV}$, from which

$$T \simeq \frac{50\,\text{MeV}}{8.617 \times 10^{-11}\,\text{MeV K}^{-1}} = 5.8 \times 10^{11}\,\text{K}$$

for the corresponding temperature in kelvin.

16.8.2 Evidence for Superfluidity in Neutron Stars

Just as for certain systems in condensed matter—though for different microscopic reasons—in neutron stars the neutrons and protons can exhibit properties of essentially

- zero resistance to mass flow *(superfluidity)*, or
- zero resistance to charge flow *(superconductivity)*.

This can have strong influence on the rotational and magnetic properties of the neutron star, as well as its rate of cooling.

For convenience we shall sometimes use "superfluidity" to mean either superfluidity or superconductivity.

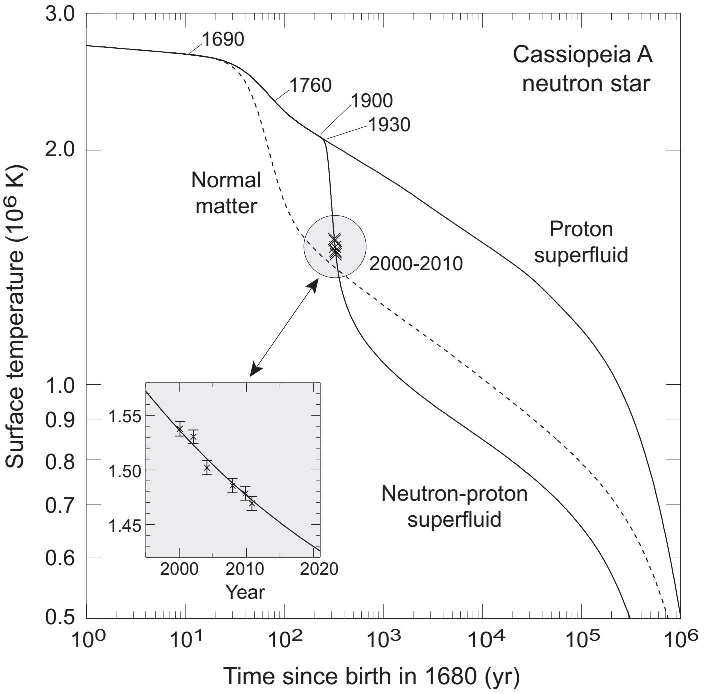
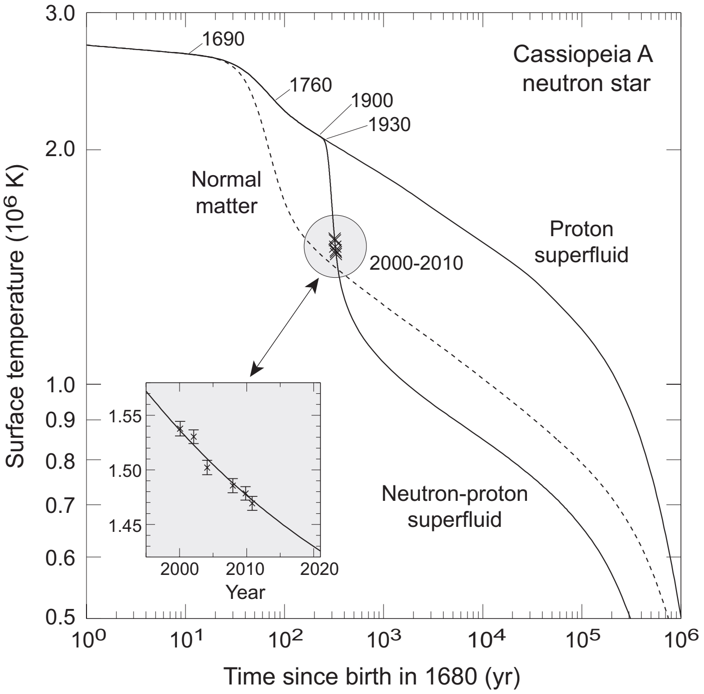


Figure 16.9: Cas A neutron star cooling. Curves indicate theory: "Normal matter" (dashed curve) assumes no superfluidity, the solid curve labeled "Proton superfluid" assumes only protons to be superfluid, and the solid curve labeled "Neutron-proton superfluid" assumes both protons and neutrons to be superfluid. Predicted temperatures are marked beginning 10 years after the birth of the neutron star in ∼1680. Chandra data points (×) suggest rapid cooling.

The Cas A Neutron Star: The Chandra X-ray Observatory discovered a compact object in the Cas A supernova remnant.

- It was subsequently identified as the neutron star left over from a supernova that occurred in the year 1681 ± 19.
- The corresponding age of about 330 years makes the Cas A neutron star the youngest known.

Evidence for superfluidity from cooling is shown in Fig. 16.9.



A Possible Superfluid Phase Transition: The theoretical curves suggest substantial differences between neutron stars with "normal" matter and those containing superfluids.

- Proton superconductivity sets in soon after formation.
  - This *suppresses neutrino emission* and
  - *lowers the cooling rate* relative to normal matter.
- When the core neutrons also become superfluid around 1930 the crust is predicted to cool very quickly for several hundred years.
- The *rapid drop of surface temperature* observed by Chandra between 1999 and 2010 has been interpreted as a *phase transition to superfluid neutrons* in the core.

16.9 Hydrostatic Equilibrium in General Relativity

The discussion of neutron stars has been based primarily on *Newtonian gravity*.

- This is adequate at a qualitative level.
- However, gravity for neutron stars is of sufficient strength that a quantitative description of them requires *general relativity (GR)*,
- with their structure determined by solving the *Einstein equations* for their dense-matter interior.
- This task is beyond our present scope. It is taken up in

  Modern General Relativity:
  Black Holes, Gravitational Waves, and Cosmology
  Mike Guidry, Cambridge University Press, 2019

  to which the reader is directed for more details.

However, let us sketch briefly how hydrostatic equilibrium is modified by general relativity in neutron stars.

16.9.1 The Oppenheimer–Volkov Equations

Stable neutron stars are in

- *hydrostatic equilibrium*, with gravity balanced against pressure-gradient forces, just as for normal stars.
- However, when gravity is derived from general relativity the corresponding equations for hydrostatic equilibrium are modified in a non-trivial way.

By assuming a perfect fluid (no shear or viscosity) the GR equations for hydrostatic equilibrium can be written in the form

$$4\pi r^2 dP(r) = \frac{-m(r)dm(r)}{r^2} \left(1+\frac{P(r)}{\varepsilon(r)}\right)\left(1+\frac{4\pi r^3 P(r)}{m(r)}\right)\left(1-\frac{2m(r)}{r}\right)^{-1}$$

$$dm(r) = 4\pi r^2 \varepsilon(r) dr.$$

where P is pressure, $\varepsilon(r)$ is energy density, and units have been chosen so that the gravitational constant G is equal to one.

- The first equation expresses hydrostatic pressure balance for a fluid in general relativity and
- the second equation implies conservation of mass–energy;
- These are termed the *Oppenheimer–Volkov* equations.

It is instructive to compare the Oppenheimer–Volkov equations with their Newtonian counterparts.

16.9.2 Comparison with Newtonian Gravity

The equations of hydrostatic equilibrium ($G = 1$ units) for *Newtonian gravity* are

$$4\pi r^2 dP(r) = -\frac{m(r)dm(r)}{r^2} \qquad dm(r) = 4\pi r^2 \rho(r)dr,$$

while the corresponding *GR equations* are

$$\begin{aligned} 4\pi r^2 dP(r) &= \frac{-m(r)dm(r)}{r^2} \\ &\quad \times \left(1 + \frac{P(r)}{\varepsilon(r)}\right)\left(1 + \frac{4\pi r^3 P(r)}{m(r)}\right)\left(1 - \frac{2m(r)}{r}\right)^{-1} \\ dm(r) &= 4\pi r^2 \varepsilon(r)dr. \end{aligned}$$

Comparing these equations indicates that

- the formulation of hydrostatic equilibrium in GR is equivalent to that in Newtonian gravity provided that
- energy density is substituted for mass density, $\rho c^2 \to \varepsilon$,
- *except for three correction factors* (in red in parentheses) in the GR version that depend on the pressure and the mass.
- These factors represent *GR corrections to Newtonian gravitation*.

Newtonian Hydrostatic Equilibrium:

$$4\pi r^2 dP(r) = -\frac{m(r)dm(r)}{r^2},$$

General Relativistic Hydrostatic Equilibrium:

$$4\pi r^2 dP(r) = \frac{-m(r)dm(r)}{r^2}$$
$$\times\left(1+\frac{P(r)}{\varepsilon(r)}\right)\left(1+\frac{4\pi r^3 P(r)}{m(r)}\right)\left(1-\frac{2m(r)}{r}\right)^{-1}.$$

- In stars described by Newtonian gravity,
  - ε is *dominated by baryon rest mass* and
  - *baryons don't contribute much to pressure* (which is dominated by electrons).
- Thus we have

  $$\frac{P(r)}{\varepsilon(r)} \sim 0 \qquad \frac{P(r)}{M(r)} \sim 0,$$

  and the first two correction factors in red are

  $$\left(1+\frac{P(r)}{\varepsilon(r)}\right) \simeq 1 \qquad \left(1+\frac{4\pi r^3 P(r)}{m(r)}\right) \simeq 1.$$

- For Newtonian gravity generally $2m(r)/r \sim 0$ and

  $$\left(1-\frac{2m(r)}{r}\right)^{-1} \simeq 1.$$

- Thus, in weak gravity *GR hydrostatic equilibrium takes the same form as Newtonian hydrostatic equilibrium*.

Newtonian Hydrostatic Equilibrium:

$$4\pi r^2 dP(r) = -\frac{m(r)dm(r)}{r^2},$$

General Relativistic Hydrostatic Equilibrium:

$$4\pi r^2 dP(r) = \frac{-m(r)dm(r)}{r^2}$$
$$\times \left(1+\frac{P(r)}{\varepsilon(r)}\right)\left(1+\frac{4\pi r^3 P(r)}{m(r)}\right)\left(1-\frac{2m(r)}{r}\right)^{-1}.$$

- Conversely, in stronger gravity the three factors in red in the Openheimer–Volkov hydrostatic equation are *all greater than one,*

$$\left(1+\frac{P(r)}{\varepsilon(r)}\right) > 1 \quad \left(1+\frac{4\pi r^3 P(r)}{m(r)}\right) > 1 \quad \left(1-\frac{2m(r)}{r}\right)^{-1} > 1$$

and these three factors

  - cause *deviations between Newtonian and GR gravity*, and
  - in general make *GR gravity stronger than Newtonian gravity*.

One of the most important consequences following from these differences between Newtonian and general relativistic gravity is that

- *in GR, gravity is stronger* and it is
- *enhanced by coupling to pressure*.
- This will imply ultimately that there are *fundamental limiting masses* for stable strongly-gravitating objects

In GR, if the mass is large enough *no amount of pressure can prevent gravitational collapse* to a black hole.

16.10 Pulsars

In 1967 something remarkable was discovered in the sky:

- a star that *appeared to be pulsing on and off* with a period of about a second.
- Quickly, even faster *"pulsars"* were discovered

and the fastest now known (the *millisecond pulsars*) pulse on and off at nearly a thousand times a second.

Pulsars exhibit several common characteristics:

1. They have well-defined periods that challenge the accuracy of the best atomic clocks.
2. The measured periods range from tens of seconds down to 1.4 milliseconds.

   > 1.4 ms corresponds to more than *700 revolutions per second*, implying a 20-km wide object spinning as fast as a kitchen blender.

3. The period of a pulsar decreases slowly with time.
   - The typical rate of decrease is a few billionths of a second each day,
   - implying that the pulsation frequency will drop to zero after about 10 million years for typical pulsars.

We shall now argue that *only a neutron star* can cause this.

16.10.1 The Pulsar Mechanism

The observational details for pulsars are inconsistent with an actual pulsation on that timescale for realistic objects.

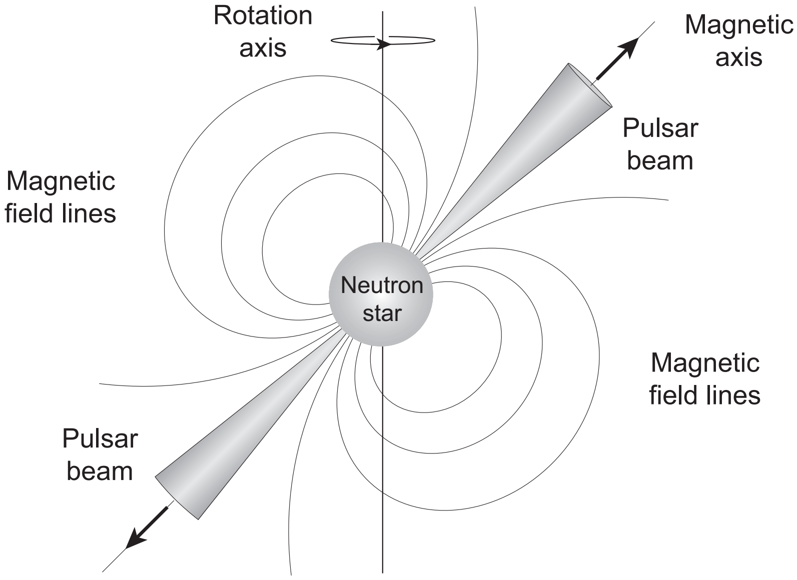
- However, a rotating star could *appear to pulse* if it had some way to emit light in a beam that rotated with the source (like a lighthouse).
- What kind of object would be consistent with observed pulsar periods?
- Simple calculations show that only a very dense object could rotate fast enough and not fly apart because of the forces associated with the rapid rotation.
- A white dwarf is not dense enough.
  - The minimum rotational period for a typical white dwarf would be several seconds;
  - for shorter periods it would fly apart.
- But a neutron star is so dense that it could rotate more than a thousand times a second and still hold together.
- This qualitative inference, augmented by much more detailed considerations, leads to the conclusion that

> The only plausible explanation is that pulsars are *rapidly spinning neutron stars,* with a mechanism to beam radiation in a *lighthouse effect.*

The Lighthouse Mechanism

A magnetic field varying in time produces an electrical field.

- Thus, the rapidly spinning magnetic field of the pulsar generates a very strong electrical field around the neutron star.
- This field accelerates electrons away from the surface at "hot spots" near the magnetic poles and these accelerated electrons produce radiation by the synchrotron effect.
- The synchrotron radiation is beamed strongly in the direction of electron motion.
- These beams rotate with the star, but the magnetic axis does not generally coincide with the rotation axis (recall Earth), so the beams rotate in a kind of corkscrew fashion:



- If these gyrating beams sweep over the Earth, they act similar to a lighthouse and we observe flashes of light.

Thus, the neutron star appears to be pulsing to an observer.

Table 16.1: Some typical magnetic field strengths

Object	Strength (gauss)
Earth's magnetic field	0.6
Simple bar magnet	100
Strongest sustained laboratory fields	4×10^5
Strongest pulsed laboratory fields	10^7
Strong magnetic stars	$10^4 - 10^5$
Radio pulsars	$10^{10} - 10^{12}$
Magnetars	10^{12}–10^{15}

16.10.2 Magnetic Fields

Some pulsars contain the *strongest magnetic fields in our galaxy*; many of their basic properties derive from these fields.

- Some typical magnetic field strengths for various objects are listed in Table 16.1 (1 tesla = 10^4 gauss).
- From the table, the two classes of objects with the largest known magnetic fields are seen to be

  1. *radio pulsars* and
  2. *magnetars* (discussed below),

  both of which involve rotating neutron stars.

Very strong magnetic fields are likely common for neutron stars but deducing that is more difficult if the neutron star is not a pulsar or magnetar.

16.10.3 The Crab Pulsar

The first pulsar was found by Jocelyn Bell and Anthony Hewish at the Cambridge radio astronomy observatory in 1967.

- The most famous pulsar was discovered shortly after that.
- It lies in the Crab Nebula (M1), which is about 7000 light years away in the constellation Taurus.
- The *Crab Pulsar* rotates about 30 times a second, emitting a double pulse in each rotation in the radio through gamma-ray spectrum.
- In visible light,
  - the Crab Pulsar appears to be a magnitude 16 star near the center of the nebula, but
  - stroboscopic techniques reveal it to be pulsing.

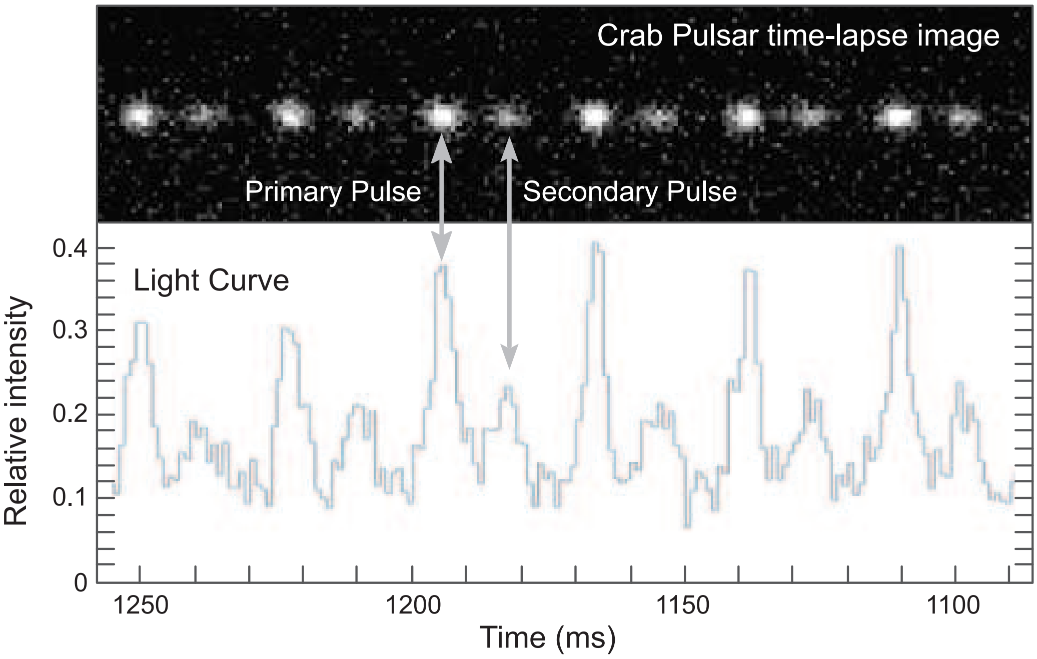
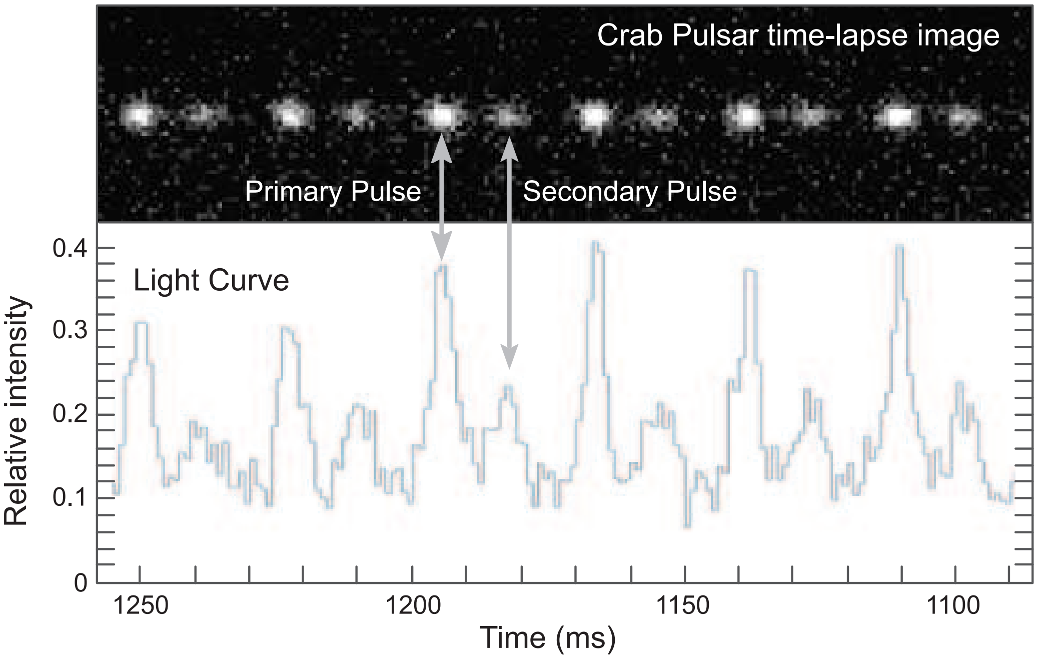


Figure 16.10: Light pulses from the Crab Pulsar. In this composite of European Southern Observatory data, the pulsar is shown in a time lapse image at the top and the light curve is displayed at the bottom on the same timescale.

Figure 16.10 shows the Crab Pulsar in action.

- The sequence is a composite of images taken through 3 different filters, all in the visible spectrum.
- Both the image sequence and the light curve show clearly the "double pulsing" of the Crab:
- in each cycle there is a strong primary pulse followed by a much weaker secondary pulse.
- The period (time between successive primary or secondary pulses) implies one primary and one secondary pulse about 30 times every second.



- This double pulsing effect can be explained by the lighthouse model if
- the geometry is such that the beam from one magnetic pole sweeps more directly over the Earth but the beam from the other pole does so only partially.
- The Crab Pulsar emits visible light (and X-rays and gamma rays), but
- most pulsars are detectable only by their radio frequency radiation.
- However, a few pulse strongly in other wavelength bands.

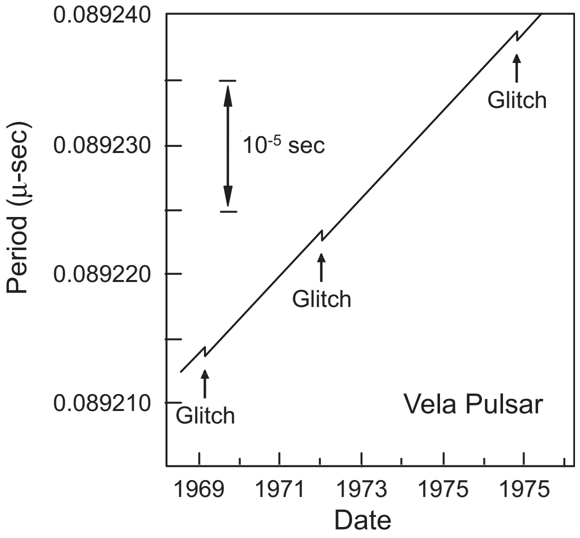


Figure 16.11: Glitches for the Vela Pulsar.

16.11 Pulsar Spindown and Glitches

In some pulsars "glitches" are observed where the spin rate suddenly jumps to a higher value (Fig. 16.11).

- The fractional change in period caused by a glitch is typically from 10^{-6} to 10^{-9} of the original period.
- Glitches indicate some internal rearrangement has altered the rotation rate by a small amount.
  - *Proposal:* "Starquakes" in the dense crust cause the neutron star to contract slightly and thus to spin faster (angular momentum conservation).
  - *Another:* Angular momentum stored in circulation of an internal superfluid liquid is suddenly transferred to the crust, altering the rotation rate.

16.12 Millisecond Pulsars

As a pulsar radiates energy away its *spin rate decreases slowly*.

- This change is small but can be measured very precisely.
- The rate of change in the rotational period for a radio pulsar can be used to estimate the *strength of the magnetic field* associated with the neutron star.
- Since pulsars are slowing down with time as they emit energy both in electromagnetic and gravitational waves, we may expect that

> *The fastest pulsars are the youngest pulsars.*

- For example, the Crab Pulsar is young (less than 1000 years), and pulses 30 times a second.
- However, *this reasoning breaks down* for those *pulsars with millisecond periods*.
- For many of these fast pulsars there is evidence that they are *old, not young* as we would expect for the fastest spin rates.
- This evidence consists primarily of
  - the rate at which the pulsar spin is slowing, and
  - where the millisecond pulsars are found.

- For example, the first millisecond pulsar discovered, PSR $1937 + 21$, is very fast but it is spinning down very slowly.

This is an example of the *standard pulsar naming system* where

- the designation PSR indicates a pulsar,
- the first part of the number gives the right ascension in hours and minutes, and
- the second part of the number gives the declination (with a sign) in degrees.

- This slow spindown rate implies that it has a weak magnetic field and is old.
- (Older pulsars should have weaker fields and these should be less effective than younger, stronger fields in braking their motion.)
- Also, many of the millisecond pulsars that have been discovered are found in globular clusters, which contain an old population of stars.
- Therefore, they are not likely to be sites of recent supernova explosions that could have produced young pulsars since core collapse supernovae occur in very short-lived, massive stars.

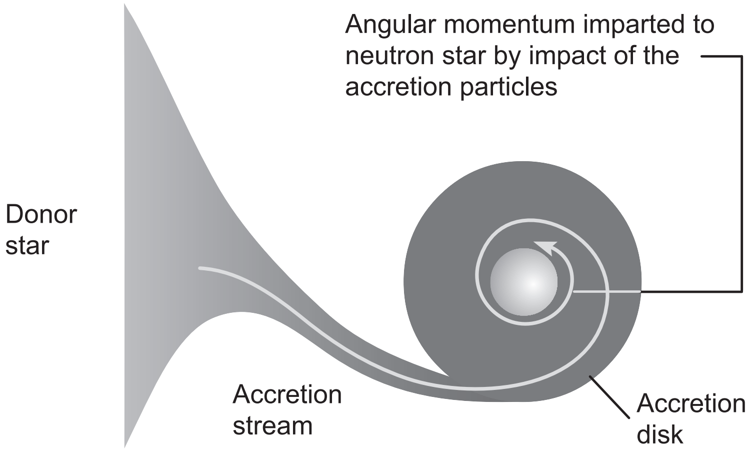

Angular momentum imparted to neutron star by impact of the accretion particles
Donor star
Accretion stream
Accretion disk

Figure 16.12: The spin-up mechanism for producing millisecond pulsars.

- The most plausible way of explaining the contradiction that the fastest pulsars seem very old is that

> *Millisecond pulsars have been "spun up".*

- The mechanism involves mass transfer in binaries that adds angular momentum to the neutron star (Fig. 16.12).
- This accretion mechanism (*binary spinup*) transfers angular momentum from orbital motion to rotation of the neutron star.
- Later, after the neutron star has been spun up to high rotational velocity, the primary star may become a supernova and disrupt the binary system.
- This leaves the rapidly spinning but old neutron star as a millisecond pulsar that defies the systematics expected from the evolution of isolated neutron stars.

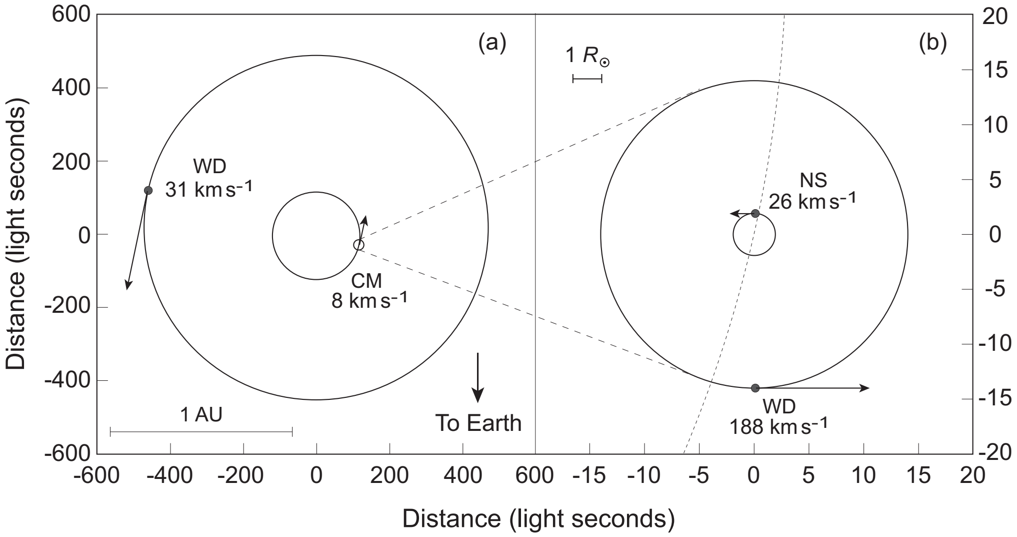


Figure 16.13: Orbits of the triplet hierarchical system PSR J0337+1715. (a) Orbits of the outer white dwarf (WD) and the center of mass (CM) for the inner white dwarf and neutron star pair. (b) Left side scaled up by a factor of 30 to show the orbits for the inner white dwarf and neutron star (NS). Arrows indicate orbital velocities for the center of mass of the inner binary and the individual white dwarfs and neutron star. All orbits lie almost in the same plane, are nearly circular, and have a tilt angle $i \sim 39°$.

The Pulsar–WD–WD Triplet PSR J0337+1715:

A rather exotic example of a millisecond pulsar is the triple-star system PSR J0337+1715, which contains

- a millisecond radio pulsar of period 2.3 ms and two white dwarfs, with orbits shown in Fig. 16.13.
- This is a *hierarchical triple-star system,* meaning that two of the stars are relatively close to each other and the other is much further away.

Such systems can have long periods of dynamical stability.

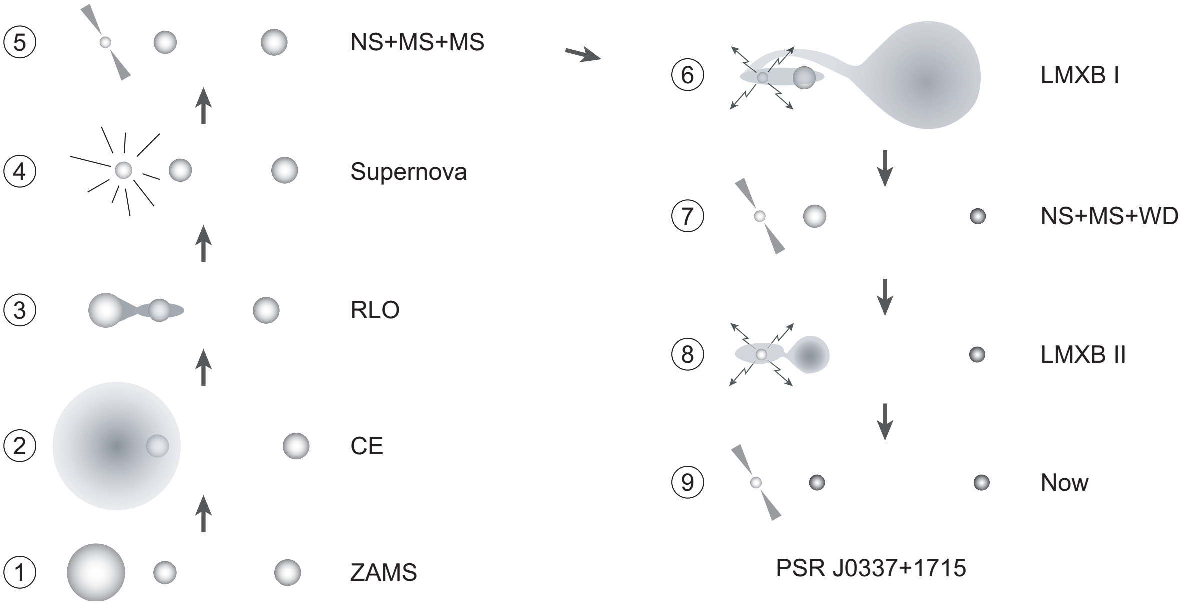
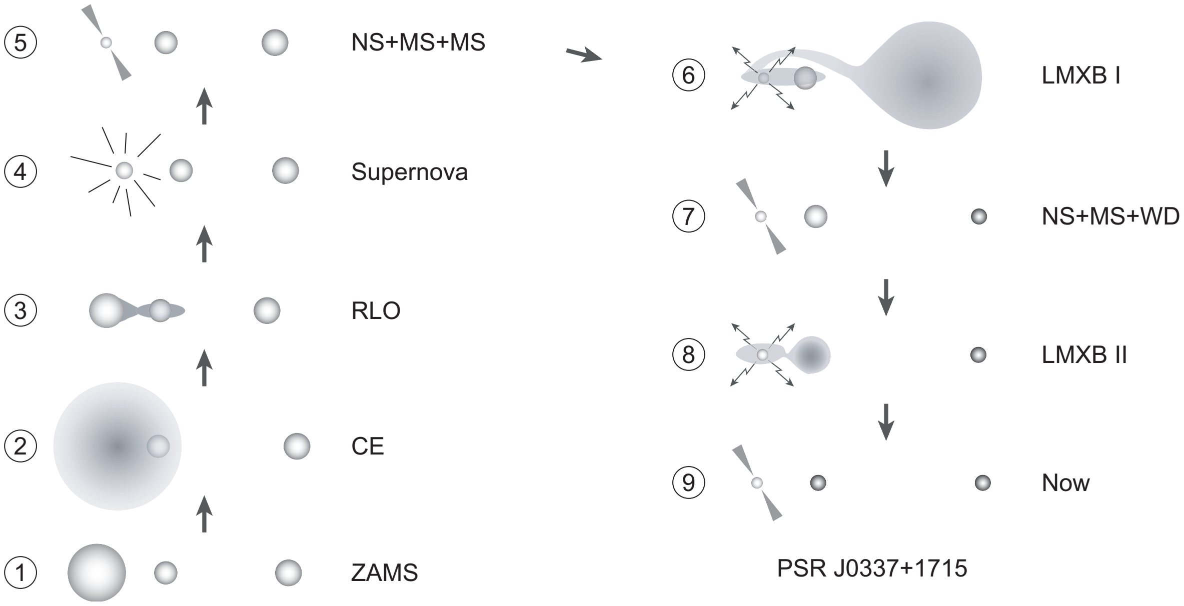


Figure 16.14: Evolutionary history of PSR J0337+1715.

PSR J0337+1715 has an *interesting evolutionary history* that is sketched in Fig. 16.14.



According to the computed scenario sketched above, this system underwent

- a *common envelope (CE) phase*,
- three periods of *Roche lobe overflow (RLO)*,
- a *supernova*, and
- Two *low-mass X-ray binary (LMXB)* episodes

to arrive at the present configuration of a neutron star + WD + WD system in which the neutron star is a *millisecond pulsar*.

- This explanation stretches current understanding of stellar evolution and stellar interactions to the limit.
- Hence PSR J0337+1715 should prove to be an excellent laboratory to study many aspects of stellar evolution, such as *common envelope phases* and *binary spinup*.

The triple-degenerate system PSR J0337+1715 also is extremely promising as a test of the *strong equivalence principle* of general relativity because of

- the *large gravitational acceleration* of the inner pulsar-WD binary by the outer white dwarf, and
- the *precise timing* afforded by the pulsar.

In this context the strong equivalence principle asserts that

- the neutron star and inner white dwarf should *fall in the same way in the gravitational field* of the outer white dwarf,
- despite their having *very different gravitational binding energies*.

Any observed deviation from this behavior would signal a *breakdown of general relativity*.

16.12.1 Binary Pulsars

Several binary star systems are known in which

- both components are neutron stars and one component is observed as a pulsar (*binary pulsars*), or
- both components are observed as pulsars (*double pulsars*).

Formation of Neutron-Star Binaries:

Binary neutron stars are of considerable interest for stellar physics because of the question of how such systems could form.

- Either a binary star system
  - survives two successive supernova explosions to form the neutron stars,
  - without disrupting the binary gravitationally, or
- Two free neutron stars in a dense cluster capture gravitationally into a binary.
- Neither scenario is easy to pull off and each appears to be possible only under very special conditions.

Nevertheless, binary neutron stars may be rare but they are observed!

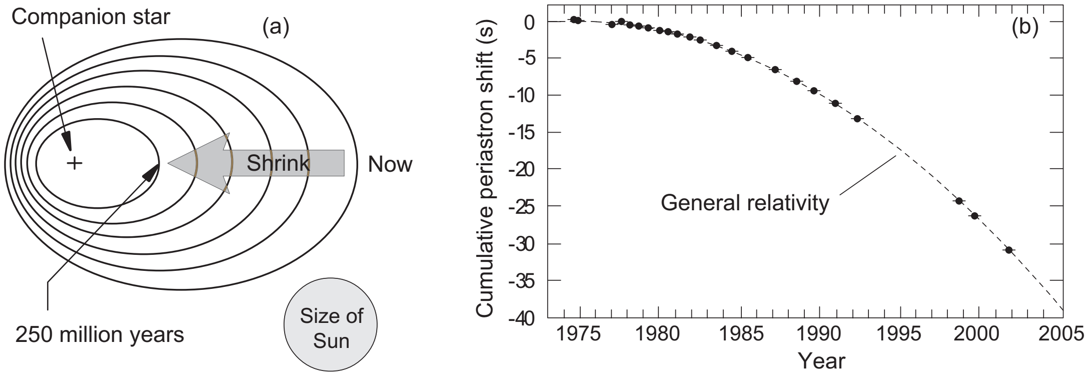


Figure 16.15: (a) Orbit of the Binary Pulsar and its decay by gravitational wave emission, drawn to scale with the Sun shown for comparison. (b) Shift of periastron time because of gravitational wave emission. Dots with error bars indicate data; the curve is the prediction of general relativity.

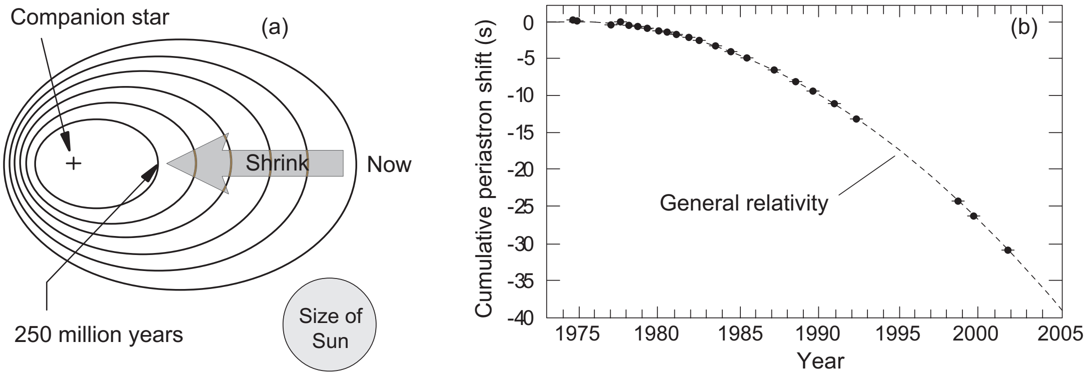
Laboratories for Testing General Relativity:

Binary pulsars and double pulsars are of great value in their own right as exotic endpoints of stellar evolution, but

- they also provide extremely precise tests of the general theory of relativity.
- This follows because pulsar periods give precise timing.
- Thus the discovery of one (better yet, two) pulsars in a binary system permits precise tests of gravitational theory.

For example, as illustrated in Fig. 16.15(a),

- the orbital semimajor axis for the *Binary Pulsar* is observed to decay by about three millimeters per revolution.
- This is the amount expected because of emission of gravitational waves predicted by general relativity.



Likewise, as illustrated in Fig. (b) above,

- the time of closest approach (periastron) between the two neutron stars has been observed to shift in precise agreement with the predictions of general relativity.
- The precise tracking of the Binary Pulsar orbit was the first compelling (although indirect) proof that gravitational waves exist.
- With the detection of a gravitational wave produced in the merger of two black holes by LIGO (Laser Interferometer Gravitational-Wave Observatory) in 2015,
- the evidence became direct that gravitational waves—the last major untested prediction of Einstein's general relativity—exist and can be observed.

This confirmation came 100 years after gravitational waves were predicted by Einstein (though Einstein had later doubts about whether they were physical).

16.13 Magnetars

Neutron stars have extremely strong magnetic fields. However, a new class of spinning neutron stars with abnormally large magnetic fields, even for a neutron star, have been discovered.

- These have been called *magnetars.*
- The magnetar SGR 1900+14 is estimated to have a magnetic field so strong ($\sim 10^{15}$ gauss) that if a magnet of comparable strength were placed halfway to the Moon, it could pull a metal pen out of your pocket on Earth!

SGR (soft gamma-ray repeater) indicates a magnetar. Like pulsars, the 1st part of the number gives the right ascension in hours and minutes, and the 2nd part gives the declination ($\pm$) in degrees.

- In these rotating neutron stars it thought that the huge magnetic fields act as a brake, slowing the rotation.
- *One proposal:* This slowing of rotation disturbs the interior structure and "starquakes" release energy that cause emission of bursts of gamma rays.

These are also called *soft gamma ray repeaters* (SGR):

- "Soft" means that the gamma rays are of low energy (in fact, they lie more in the X-ray portion of the spectrum);
- "Repeater" means that the bursts can repeat.

Chapter 17

Black Holes

We have seen that the endpoints for stellar evolution grow increasingly bizarre as the mass of a star is increased.

- For lighter stars the final chapter is a *white dwarf* of incredible density, stabilized by electron degeneracy.
- For more massive stars the endpoint is a *neutron star,* with a density exceeding that of atomic nuclei, stabilized by neutron degeneracy pressure.
- In this chapter we consider the strangest endpoint of all:
  - The most massive stars collapse until the mass is concentrated at a point singularity.
  - The singularity is surrounded by a one-way spacetime membrane called the *event horizon* that forbids the escape of light or matter.

This most extreme consequence of modern gravitational physics is called a *black hole.*

For even a basic understanding of black holes, *general relativity (GR)* is essential.

- A systematic introduction to general relativity is outside the present agenda. For a comprehensive introduction, see:

Modern General Relativity:
Black Holes, Gravitational Waves, and Cosmology
Mike Guidry, Cambridge University Press, 2019

- But in this chapter some essential concepts and a few formulas will be imported from that book to allow a meaningful discussion of collapse to a black hole.

Some observational evidence for black holes will then be discussed for two categories:

- *High-mass compact objects* in spectroscopic binary systems.
- *Gravitational waves* generated by the interaction of black holes with other black holes or with neutron stars.

17.1 The Failure of Newtonian Gravity

The Newtonian theory of gravity is a remarkably good description of the Universe.

- It gives predictions for most phenomena that are in practically exact agreement with observations (and the corresponding predictions of general relativity).
- However, there is a small set of phenomena for which general relativity gives the correct prediction but Newtonian gravity fails.

These failures of Newtonian gravity typically share some combination of three characteristics:

1. Gravity becomes extremely strong, by measures that we shall quantify shortly.
2. Characteristic velocities approach the speed of light.
3. A measurement may require sufficient precision that even small deviations from Newtonian gravity become manifest.

> An example of the latter is the *Global Positioning System (GPS),* where the precise timing required to determine position implies that even the special and general relativistic corrections for low velocity in Earth's weak gravitational field are large.

If any of these conditions is fulfilled, the predictions of Newtonian gravity begin to fail. In the extreme case where all are true general relativity becomes the only viable theory of gravity.

- Black holes tend to fall into this latter category.
- Newtonian concepts often are unreliable or even in downright error where the physics of black holes is concerned.

17.2 General Covariance

The essential idea of both special and general relativity is an extremely powerful principle:

> The laws of physics should not depend on the reference frame in which they are formulated and so should be unchanged by transformation to a new coordinate system.

The basic difference between special and general relativity then is just that

- In general relativity the laws must be invariant under the most general possible transformations between coordinate systems.
- Special relativity requires only invariance only under a more restricted set of transformations (between *inertial frames:* coordinate systems that are not accelerated with respect to each other)

17.2.1 The Principle of Equivalence

The fundamental insight that allowed Einstein to generalize special relativity to a theory of gravity began with the idea—known since the time of Galileo—that

- objects of different mass fall at the same rate in a gravitational field.
- This is one formulation of the *(weak) equivalence principle.*

An alternative formulation is that

- the *inertial mass* of an object, corresponding to the mass m in Newton's second law, $F = ma$,
- is measured to be equivalent to the *gravitational mass* of that same object, corresponding to the mass m in the gravitational law $F = GmM/r^2$.

Starting from this insight, Einstein was led to propose that

- it is impossible locally to distinguish the effect of gravity from the effect of an arbitrary acceleration.
- This is called the *(strong) equivalence principle*, which henceforth will be termed simply the *equivalence principle.*

Furthermore, Einstein reasoned that

- since the acceleration of an object by gravity was independent of the mass or any other characteristic of the object,
- the effect of gravity *cannot be a property of objects in spacetime* and therefore must be a property of spacetime itself.

This led Einstein eventually to the central thesis of general relativity:

> Spacetime is *curved*, and gravity is *not a force* but rather corresponds to the motion of *free particles* in a *curved spacetime*.

In this view the Earth is in orbit around the Sun,

- not because of a force acting between them, but
- because the gravitational field of the Sun curves the spacetime around it and
- the Earth follows freely a curved path in that curved spacetime.

> This means that general relativity is a theory about the *geometry of spacetime.*

In relativity it is natural to unite space and time in a 4-dimensional manifold called *spacetime.*

- The coordinates of a point P in spacetime are given by the *4-vector* x^μ denoted by

  $$x^\mu \equiv (x^0, x^2, x^3, x^4) = (ct, x, y, z) \qquad (\mu = 0, 1, 2, 3),$$

  where x, y, and z are spatial coordinates, t is the time, and c is the speed of light.

- Since general relativity is invariant even under transformations between non-inertial frames, it can describe gravity.

The most powerful and useful mathematical implementation of general relativity is in terms of objects called *rank-n tensors.*

- A tensor of rank n may be viewed mathematically as a function (map) of n vectors into the real numbers.
- This implies that components of tensors evaluated in some basis carry a total of n upper and lower indices, and transform in a particular way under coordinate transformations.

Practically, tensors are

- an extension of vectors to objects that *generalize the vector transformation law* and that
- may carry *more than one index* when evaluated in a basis.

Indeed, a vector may be viewed as a rank-1 tensor.

17.3 The Geometry of Spacetime

The geometry of spacetime is described by a rank-2 tensor called the *metric tensor,* $g_{\mu\nu}$.

- Einstein showed that $g_{\mu\nu}$ can be viewed as the *source of the gravitational field.*
- Thus the problem in general relativity is “simple”:

> Just determine the metric tensor for the manifold, which then determines the complete effect of gravity.

- But not so fast! Not only does the gravitational “force” acting on mass and energy in spacetime result from the curvature of spacetime, but that same mass and energy acts on spacetime to curve it.
- This implies that general relativity is a *highly non-linear theory* (to determine the metric you must already know the metric).

> Thus the equations of general relativity
>
> - can be *written concisely using tensors*, but
> - they are *extremely difficult to solve*.

17.4 Curvature and the Strength of Gravity

Our primary concern here will be with *strong gravity.*

- But what does that mean in this context?
- Gravity in GR is a property of *curved spacetime.*
- This causes the path of light to be *bent in a gravitational field*.

Thus a natural measure of the strength of the gravitational field near a spherical object is the ratio of the curvature of space to the curvature of the surface of that object.

In general relativity light follows a curved path in a gravitational field.

- A *radius of gravitational curvature* $r_{\rm c}$ may be obtained by fitting a circle to the local curved path.
- This gives $r_{\rm c} = c^2/g$, where g is the gravitational acceleration and c the speed of light.
- A natural measure of gravitational strength at the surface of a spherical object such as a star is then

$$\frac{R}{r_{\rm c}} = \frac{\text{Actual radius}}{\text{Light curvature radius}} = \frac{GM}{Rc^2},$$

  where $g = GM/R^2$ was used.
- Then weak gravity is characterized by $GM/Rc^2 \ll 1$, but if $R/r_{\rm c} \sim 1$ a gravitational field may be characterized as strong.

It is also instructive to multiply the above equation by m/m and write it in the form

$$\frac{R}{r_{\rm c}} = \frac{GMm/R}{mc^2} = \frac{E_{\rm g}}{E_0} = \frac{\text{Gravitational energy}}{\text{Rest mass energy}}.$$

- Thus, the weak-gravity condition $GM/Rc^2 \ll 1$ implies that *the gravitational energy of a test particle is much less than its rest mass energy*.

If the gravitational field is strong by this natural standard, general relativity must be used to describe gravity.

Table 17.1: Gravitational strengths $R/r_{\rm c}$ at the surface of some objects

Object	R(km)	M(kg)	$\rho\,(\mathrm{g\,cm^{-3}})$	$g\,(\mathrm{m\,s^{-2}})$	$r_{\rm c}$(km)	$R/r_{\rm c}$
Earth	6378	6×10^{24}	5.6	9.8	9.2×10^{12}	6.9×10^{-10}
White dwarf	5500	2.1×10^{30}	$\sim10^{6}$	4.6×10^{6}	1.9×10^{7}	2.8×10^{-4}
Neutron star	10	2×10^{30}	$\sim10^{14}$	1.3×10^{12}	67.5	0.15

Most gravitational fields are weak by the natural measure

$$\frac{R}{r_{\rm c}}=\frac{\text{Actual radius}}{\text{Light curvature radius}}=\frac{GM}{Rc^2}.$$

Table 17.1 gives some examples.

- You may tend to think of Earth's gravity as relatively strong when climbing stairs, but it corresponds to a paltry $R/r_{\rm c}\sim10^{-9}$!
- Even a white dwarf has only $R/r_{\rm c}\simeq10^{-4}$.
- This is still weak on the natural scale set by light curvature (though enormous by Earth standards).
- Thus Newtonian gravity is still a rather good approximation for white dwarfs.
- But for gravity at the surface of a neutron star or the event horizon of a black hole, the gravitational curvature radius and actual radius will be comparable.

Thus a correct description of gravity for neutron stars and black holes requires general relativity.

17.5 Some Important General-Relativistic Solutions

In general relativity the rank-2 metric tensor $g_{\mu\nu}$ is both

- the source of the gravitational field and
- the description of the geometry of spacetime.
- Thus the task is to determine $g_{\mu\nu}$, which is generally dependent on the spacetime coordinates, for a given situation.

But this is a quite non-trivial task.

- In Newtonian physics the metric is fixed and specified implicitly at the beginning of a problem.
- It corresponds to
  - the flat (euclidean) spatial coordinates and
  - the time, which is assumed in Newtonian physics to be defined globally and thus the same for all observers.
- In contrast, in general relativity the metric is not known beforehand: it is *the solution of the problem*.

Thus the framework of spacetime in which the problem is formulated is itself unknown at the beginning for a GR problem.

17.5.1 The Einstein Equation

This highly-nonlinear problem can be solved because it can be shown that solutions obey the *Einstein equation*,

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = \frac{8\pi G}{c^4} T_{\mu\nu}.$$

In this expression

- The indices μ and ν each range over the labels for the spacetime dimensions $(0,1,2,3)$.
- $R_{\mu\nu}$ and R are rank-2 and rank-0 tensors called the *Ricci tensor* and the *Ricci scalar,* respectively.
- The Ricci tensor and Ricci scalar depend on the metric tensor $g_{\mu\nu}$ and describe the *curvature of spacetime*.
- $T_{\mu\nu}$ is a rank-2 tensor called the *stress–energy tensor*.
- $T_{\mu\nu}$ describes the coupling of gravity to matter, energy, and momentum.

Because of the indices, each term in the Einstein equation can be viewed as a *matrix with 16 components*.

- However *only 10 are independent* because all terms are symmetric under exchange of indices.
- Hence this deceptively simple expression actually represents *10 coupled, non-linear, partial differential equations* that determine the effect of gravity.

In the general case analytical solutions of the Einstein equation are hopeless.

- However, in some cases of physical interest the problem has a high degree of symmetry.
- This may reduce the problem to solving a much smaller set of equations that is still often formidable, but tractable.

Often only the gravitational solution outside some mass responsible for producing the gravitational field is of physical interest.

- Then if the exterior region is assumed to be a vacuum, the Einstein equation reduces to

$$R_{\mu\nu} = 0.$$

  which is called the *vacuum Einstein equation.*

- Don't be fooled by the seeming triviality of this equation either!

Because of the nonlinearity and the tensor indices, the vacuum Einstein equation is also extremely difficult to solve.

17.5.2 Line Elements and Metrics

In the following some solutions of the Einstein equation will be quoted without derivation.

- Such solutions are often called *“spacetimes”*.
- Instead of giving the metric tensor that corresponds to the solution it is common to express solutions in terms of the *line element* ds^2 ,
- where a standard notation $ds^2 \equiv (ds)^2$ has been used.
- This is related to the metric tensor $g_{\mu\nu}$ by

$$ds^2 = \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} g_{\mu\nu} dx^\mu dx^\nu \equiv g_{\mu\nu} dx^\mu dx^\nu ,$$

- where we’ve introduced in the last step the *Einstein summation convention*:

> An index repeated twice on one side, once in a lower and once in an upper position, implies a summation on that index.

- Whether a tensor index is in an upper or lower position is mathematically and physically important.
- But it will be sufficient for present purposes just to remember that, without going into details of why.

Often the line element ds^2 is just called “the metric”.

17.5.3 Minkowski Spacetime

Let’s warm up with the “trivial” case.

- The simplest possibility is no gravitational fields, so that spacetime has no curvature (flat spacetime).
- Then general relativity reduces to special relativity.
- The resulting 4-dimensional manifold is called *Minkowski spacetime*, or just *Minkowski space*.
- The corresponding metric is

$$ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2.$$

- The time-like component c^2dt^2 has a sign *opposite* that of the three space-like components dx^2, dy^2, and dz^2.
- A metric for which the terms in the line element do not all have the same sign is called *indefinite*.
- Indefinite metrics are *characteristic of physical spacetime,* whether gravity is present or not.

Thus 4-dimensional Minkowski space has

- a very different geometry than 4-dimensional Euclidean space,
- even though *both are flat* and both correspond to *spaces with no intrinsic curvature*.

In fact, the relative negative sign between space and time coordinates in the Minkowski metric *(indefinite metric)* is the source of all the "strange" behavior associated with special relativity:

- space contraction,
- time dilation,
- relativity of simultaneity,
- the "twin paradox",

all derive from the indefinite Minkowski metric.

The metric must be used to compute physical observables, which illustrates another fundamental difference between relativity and Newtonian physics.

- In a Newtonian description coordinates may be themselves physical quantities.
- For example, the value of r in spherical coordinates is a distance that could be measured.
- In general (and special) relativity, *space and time coordinates are just labels,* without direct physical significance.
- Physical quantities must be
  - computed using the metric.
  - They generally are not given directly by values of coordinates.
- This is illustrated by the metric itself:

  - $(ds^2)^{1/2}$ measures the physical length of an infinitesimal line segment.
  - By inspection this distance is not given directly by any of the coordinates.
  - Rather it is a mixture of contributions from space and time coordinates.

Example: Consider the time coordinate t.

- In Newtonian theories the *coordinate time* t is a direct measure for all observers of the passage of time.
- In Minkowski space the *proper time* τ is defined to be the time measured by a clock carried by an observer in the observers's inertial frame.
- The proper time and distance interval are related (Exercise) by $d\tau^2 = -ds^2/c^2$.
- Then from the line element ds^2,

$$d\tau = \left(1 - \frac{v^2}{c^2}\right)^{1/2} dt\,.$$

- The proper time that elapses between coordinate times t_1 and t_2 is then

$$\tau_{12} = \int_{t_1}^{t_2} \left(1 - \frac{v^2}{c^2}\right)^{1/2} dt\,.$$

- For constant velocity, this yields

$$\Delta\tau = \left(1 - \frac{v^2}{c^2}\right)^{1/2} \Delta t\,,$$

  which is just the time dilation equation of special relativity: The proper time interval $\Delta\tau$ is shorter than the coordinate time interval Δt because the square root is always less than one.

Thus time dilation in special relativity is a direct consequence of the *indefinite Minkowski metric*,

$$ds^2 = -c^2 dt^2 + dx^2 + dy^2 + dz^2.$$

- Specifically, time dilation follows from the difference in signs between the timelike and spacelike components of the metric.
- In a similar manner the Minkowski metric may be used to derive the space contraction effect and other standard features of special relativity.

17.5.4 Schwarzschild Spacetime

If gravitational fields are present the Minkowski metric no longer applies. The simplest solution in that case is obtained by assuming the spacetime where the solution is valid to be

- devoid of matter, pressure, and fields,
- independent of time, and
- spherically symmetric in the spatial coordinates.

That is, some time-independent, spherical distribution of mass is assumed to produce a gravitational field, but

- the Schwarzschild solution is valid only *outside* the mass distribution responsible for the field.
- For a spherical star, this solution would be valid beyond the radius of the star.
- For the spherical black holes to be discussed below, all the mass that is the source of the gravitational field has been crushed into a singularity at the center of the black hole.

This solution of the vacuum Einstein equation is called the *Schwarzschild spacetime,* and has the metric

$$ds^2 = -\left(1-\frac{2M}{r}\right)dt^2 + \left(1-\frac{2M}{r}\right)^{-1}dr^2 + r^2 d\theta^2 + r^2 \sin^2\theta\, d\varphi^2,$$

where

- t is a time coordinate,
- r is a radial coordinate,
- θ and φ are the usual spherical angular coordinates,
- M is the single parameter of the theory,
- In the weak-field limit M may be interpreted as the mass responsible for the gravitational field.

In this equation another standard convention of the relativity formalism has been introduced:

- A special set of units is used where $G = c = 1$.
- Thus G and c do not appear explicitly in the equations.

As for Minkowski space,

- The coordinates (t, r, θ, φ) are just labels.
- Physical quantities must be computed from the metric.

The Schwarzschild solution implies a very unusual situation if the mass M is compressed into a region smaller than the *Schwarzschild radius* r_{s} defined by

$$r_{\text{s}} = \frac{2GM}{c^2},$$

where the c and G factors have been reinserted.

- By computing observables using the metric, it is found that in this case the radius r_{s} defines an *event horizon*.
- As a consequence of the extreme curvature of spacetime at the event horizon,
  - matter or light can fall through the horizon but
  - once inside *nothing can escape,* not even light.

This solution is the simplest example of a *black hole.*

17.5.5 Kerr Spacetime

The Schwarzschild black hole described above is spherically symmetric and has no angular momentum.

- It is useful to illustrate the idea of black holes.
- However, black holes formed from gravitational collapse of stars are expected to have angular momentum.
- The solution giving black holes that are deformed and spinning is called the *Kerr spacetime.*
- It is specified in terms of what are called *Boyer–Lindquist coordinates* (t, r, θ, φ) by the metric

$$ds^2 = -\left(1 - \frac{2Mr}{\rho^2}\right) dt^2 - \frac{4Mra\sin^2\theta}{\rho^2} d\varphi dt$$

$$+\frac{\rho^2}{\Delta} dr^2 + \rho^2 d\theta^2 + \left(r^2 + a^2 + \frac{2Mra^2\sin^2\theta}{\rho^2}\right)\sin^2\theta d\varphi^2$$

with the definitions

$$a \equiv J/M \qquad \rho^2 \equiv r^2 + a^2\cos^2\theta \qquad \Delta \equiv r^2 - 2Mr + a^2.$$

- This gives a 2-parameter family of solutions in terms of the parameters a (or equivalently J) and M, where
- in the weak-field limit J may be interpreted as angular momentum and M as the mass.

Coordinates are labels without direct physical significance and the metric must be used to calculate observables. The most important features of Kerr black holes for us are:

- A Kerr spacetime has a region near the black hole but outside its event horizon called the *ergosphere*.
- A particle could enter the ergosphere and still escape, carrying off part of the rotational angular momentum and rotational energy of the black hole.
- If the rotational energy and angular momentum are removed completely from a Kerr black hole, what remains is a Schwarzschild black hole, from which no additional mass or energy can be removed.
- The spinning black hole drags the surrounding spacetime as it rotates. This is called *frame dragging*.
- Thus objects near the black hole will be dragged with the rotation of the black hole *even if no angular force acts between the object and the black hole.*
- There is a maximum possible angular momentum for a Kerr black hole of mass M that is given by

  $$J_{\text{max}} = M^2.$$

  Kerr black holes having $J = J_{\text{max}}$ are called *extremal Kerr black holes.*
- It is expected that near-extremal Kerr black holes could be relatively common.

- Schwarzschild black holes are a special case of Kerr black holes corresponding to $J = 0$.
- In principle black holes could be electrically-charged, which corresponds to yet other solutions of the Einstein equations.
- However, it is generally thought that any black holes formed in realistic astrophysical processes would be quickly charge-neutralized.
- Hence our interest here will be solely in uncharged black holes.
- It may be assumed that any real black holes are Kerr black holes, usually with $J \neq 0$.

17.6 Evidence for Black Holes

With an understanding that

- black holes are intrinsically objects that must be described by general relativity, and
- armed with a qualitative understanding of concepts from general relativity,

let us now summarize some of the observational evidence supporting the thesis that black holes exist.

- Their very name suggests that they are difficult to observe directly, but
  - if black holes are not isolated they should often be accreting matter and interacting gravitationally with nearby masses.
  - These could have observable consequences.

There are in fact strong reasons to believe in the reality of black holes, based on three kinds of observations.

1. *Massive unseen companions* in binary star systems that are strong X-ray sources.
2. *Detection of gravitational waves*, for which the properties of the wave suggest that it originated in the merger of two black holes.
3. Observational *anomalies in the centers of many galaxies*, where

   - very large masses (millions to billions of solar masses) inferred from star velocities exist,
   - often accompanied by evidence for enormous energy generation in the core of the galaxy.

Our primary interest here is in black holes with masses comparable to those of stars that are potential endpoints for stellar evolution (*stellar black holes*).

- So let us concentrate on evidence for stellar black holes in categories 1 and 2.

- At the end of the discussion some evidence for the supermassive black holes in category 3 will be presented for completeness.

17.6.1 Masses for Compact Objects in X-Ray Binaries

There is appreciable indirect evidence for stellar black holes with masses $\sim 5 - 50 M_\odot$.

- Much of this evidence comes from observation of X-ray sources powered by accretion in binary star systems.
- Most such systems are *spectroscopic binaries,* where an unseen compact object (usually a neutron star or black hole) is inferred from periodic Doppler shifts of spectral lines for the visible star.
- Typically X-ray emission in a spectroscopic binary is caused by significant accretion onto the compact companion.
- This implies a relatively small separation between components of the binary.
- Tidal interactions in close binaries tend to circularize elliptical orbits.
- Hence our discussion will be considerably simplified but not seriously compromised by assuming circular orbits.

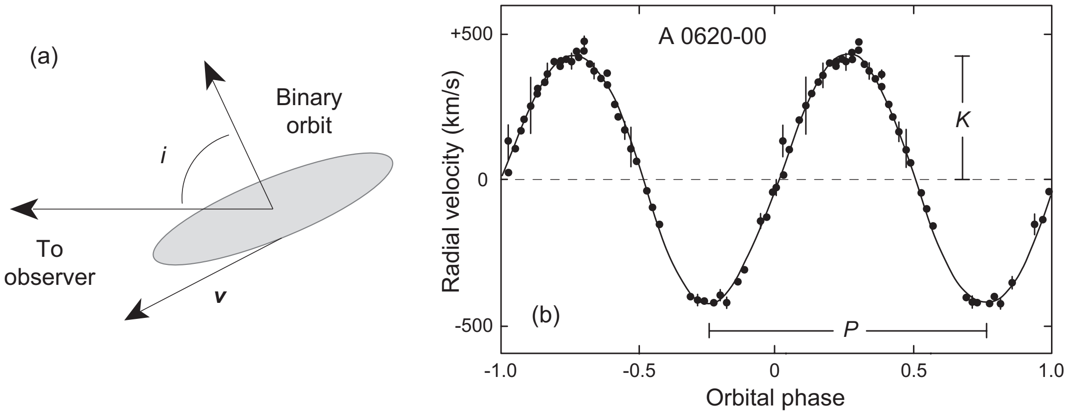


Figure 17.1: (a) Tilt angle i for a binary orbit. (b) Radial velocity curve for the spectroscopic binary A 0620–00. The period is $P = 0.323$ days and the semiamplitude is $K = 433 \pm 3\,\text{km}\,\text{s}^{-1}$.

The *mass function* $f(M)$ may be related to an observed radial velocity curve as in Fig. 17.1(b) through

$$f(M) \equiv \frac{(M\sin i)^3}{(M+M_\text{c})^2} = \frac{M\sin^3 i}{(1+q)^2} = \frac{PK^3}{2\pi G},$$

where

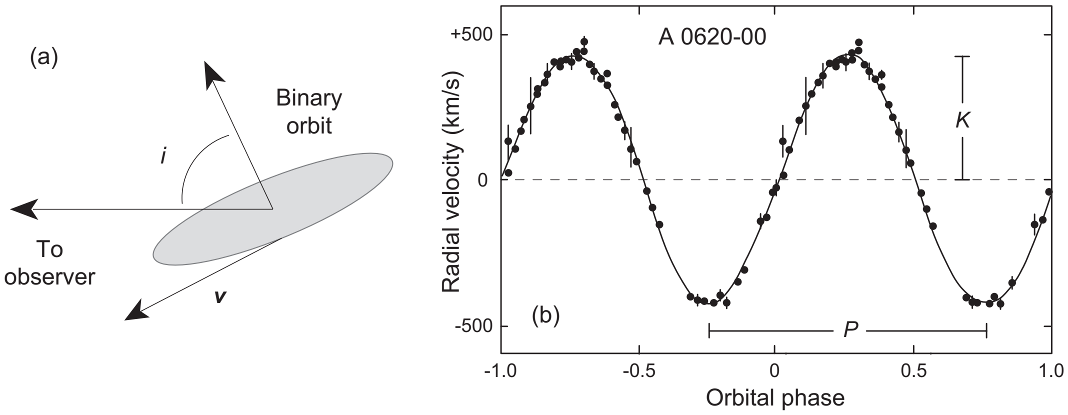
- K is the semiamplitude and P the period of the radial velocity curve,
- i is the tilt angle relative to the observer of the orbit,
- M_c is the mass of the visible companion star,
- M is the mass of the unseen component, and
- the *mass ratio* $q \equiv M_\text{c}/M$ has been introduced.

The mass function

$$f(M) \equiv \frac{(M\sin i)^3}{(M+M_{\rm c})^2} = \frac{M\sin^3 i}{(1+q)^2} = \frac{PK^3}{2\pi G},$$

is useful because

- the right side is determined by direct observation of the radial velocity curve and
- the left side is a function of the masses,
- so the measured velocity curve can be related to the masses in the binary.
- As will be seen below, the mass function *places a lower limit on the sum of the masses* in the binary.
- With some additional information can often place constraints on the mass of the unseen compact object.



The tilt angle i is illustrated in Fig. (a) above and a typical observed velocity curve for a binary system is shown in Fig. (b) above.

- The angle i is generally not known for a spectroscopic binary
- (except that the presence or absence of eclipses can place some limits on it).
- Hence the measured mass function places a *lower limit* on the mass of the unseen component if the mass of the (visible) companion can be determined.
- This is often possible from spectral systematics for the companion.

Example: Let’s compute the mass function for a binary having

- a period of $P = 5.6$ days and
- a semiamplitude for the radial velocity curve $K = 75\,\mathrm{km\,s^{-1}}$.

From the mass function equation expressed in convenient units

$$f(M) = \frac{PK^3}{2\pi G} = 1.036\times 10^{-7}\left(\frac{P}{1\ \mathrm{day}}\right)\left(\frac{K}{\mathrm{km\,s^{-1}}}\right)^3 M_\odot.$$

Inserting $P = 5.6$ day and $K = 75\,\mathrm{km\,s^{-1}}$ gives

$$f(M) = 0.245$$

for the mass function.

Suppose that the period P and velocity semiamplitude K have been determined from the observed velocity curve for a spectroscopic binary and that the quantity

$$F = F(P,K) \equiv \frac{PK^3}{2\pi G}$$

has been computed from that information. Then from

$$f(M) \equiv \frac{(M\sin i)^3}{(M+M_{\rm c})^2} = \frac{M\sin^3 i}{(1+q)^2} = \frac{PK^3}{2\pi G},$$

the unknown compact-object mass M is determined by

$$\frac{M^3\sin^3 i}{(M+M_{\rm c})^2} = F,$$

for which the solution of physical interest is given by the real root

$$M(F,M_{\rm c},i) = \left(R+\sqrt{Q^3+R^2}\right)^{1/3} + \left(R-\sqrt{Q^3+R^2}\right)^{1/3} - \frac{a}{3}$$

$$R \equiv \tfrac{1}{54}(9ab-27c-2a^3) \qquad Q \equiv \tfrac{1}{9}(3b-a^2)$$

$$a = -\frac{F}{\sin^3 i} \qquad b = -\frac{2FM_{\rm c}}{\sin^3 i} \qquad c = -\frac{FM_{\rm c}^2}{\sin^3 i} \cdot$$

Since F is known, the mass M of the compact unseen component is a function of two unknowns:

- the *mass of the visible companion* $M_{\rm c}$ and
- the *tilt angle* i.

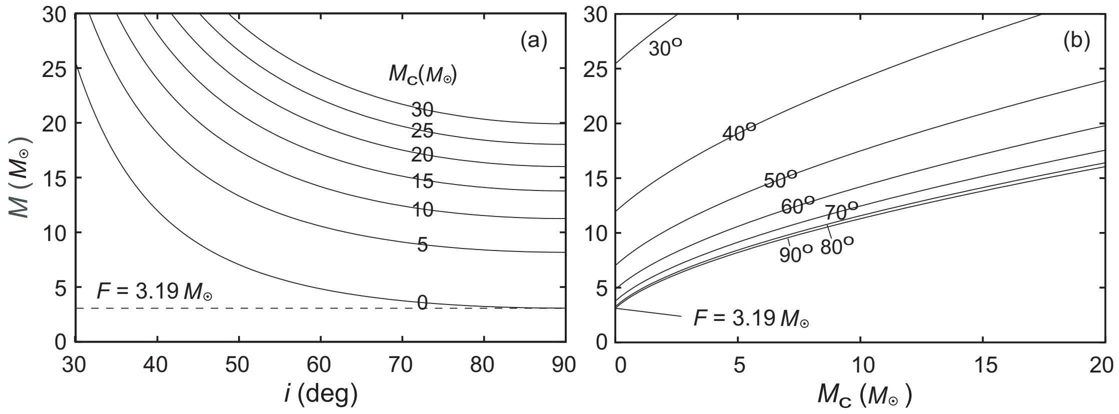


Figure 17.2: Mass plots assuming a measured mass function $F = 3.19 M_\odot$. (a) Mass M of the unseen component versus the tilt angle i for different values of the companion mass M_c. (b) Mass M of the unseen component versus M_c for different values of i. The measured mass function F is seen to set a lower limit on the unseen mass M.

In Fig. 17.2, the solution $M(F, M_c, i)$ is plotted as a function of i and M_c assuming that $F = 3.19\ M_\odot$. These figures illustrate clearly

1. The degeneracy of the unknown mass M with respect to the parameters i and M_c
2. That the measured value $F = 3.19 M_\odot$ is the minimum possible mass for the unseen component.

Point 2 already is a powerful constraint but a more precise statement about M is possible if further information can be obtained about M_c and i, as we shall demonstrate below.

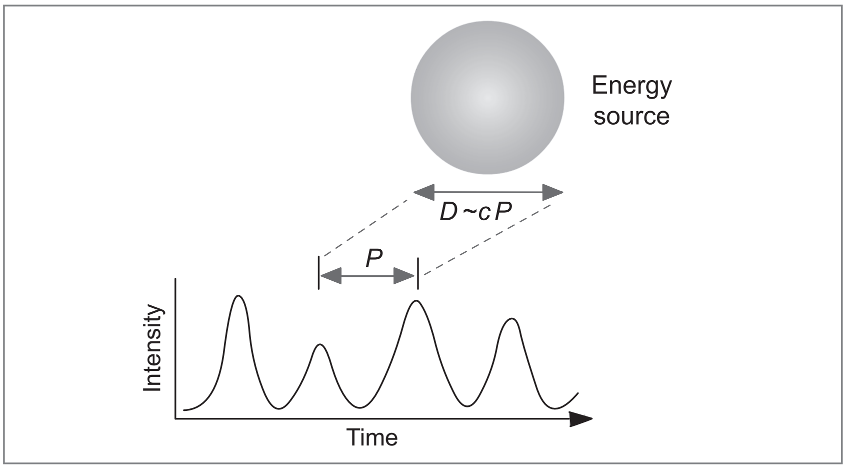
17.6.2 Causality Constraints

Another important argument that can be marshaled to determine whether a spectroscopic X-ray binary harbors a black hole is *causality*.

- If the X-ray source is observed to vary periodically, some signal must correlate the periodic variation and it cannot travel faster than light.
- Hence the maximum size of the source is limited by the finite speed of light.

If such considerations point to a very small energy source, typically a black hole or a neutron star is implicated.

If the luminosity of an energy source is periodic, some signal must tell the source to vary. The maximum size D of an object varying with a period P is the distance that light could have traveled during that time, $D \sim cP$:



The distances covered by light for various fixed times are summarized in the following table.

Time	km	AU	Parsecs
Year	9.46×10^{12}	63,240	3.07×10^{-1}
Month	7.88×10^{11}	5270	2.58×10^{-2}
Week	1.82×10^{11}	1216	5.90×10^{-3}
Day	2.59×10^{10}	173	8.41×10^{-4}
Hour	1.08×10^{9}	7.21	3.50×10^{-5}
Minute	1.80×10^{7}	0.120	5.84×10^{-7}
Second	3.00×10^{5}	0.002	9.73×10^{-9}
Millisecond	3.00×10^{2}	0.000002	9.73×10^{-12}

This causality argument places only an *upper limit* on source size and the energy-producing region may be *smaller* than the limit imposed by c. But it is a very powerful argument because it depends only on causality.

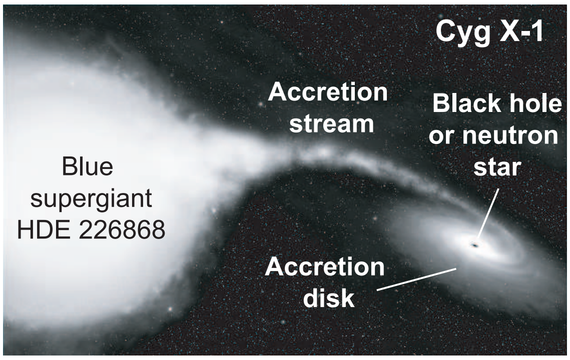


Figure 17.3: Artist's conception of the high-mass X-ray binary, Cyg X-1.

17.6.3 The Black Hole Candidate Cygnus X-1

Let us use the preceding ideas to analyze the black hole candidate Cygnus X-1.

- Optical, X-ray, and RF observations in the 1960s and 1970s determined that Cygnus X-1 is an X-ray source in a binary system consisting of
  - the visible blue supergiant HDE 226868 and
  - an unseen companion that must be a white dwarf, neutron star, or black hole
- The X-ray source flickers with a period of ms.
- From causality, this suggests that the source size is no more than a few hundred kilometers
- This rules out a white dwarf and implicates either a neutron star or a black hole.

An artist's conception is displayed in Fig. 17.3.

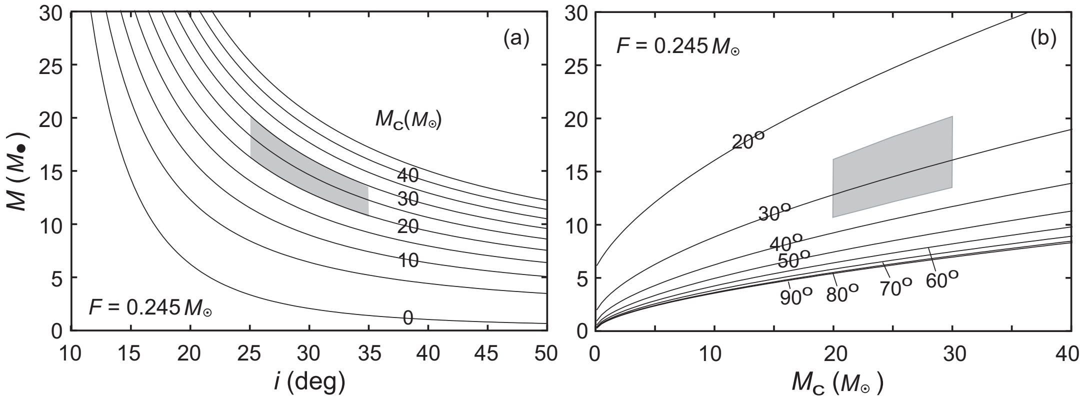


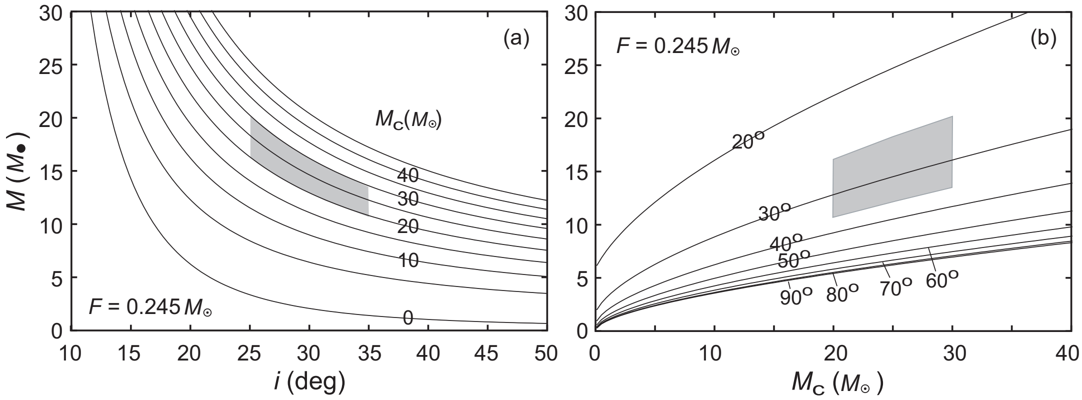
Figure 17.4: Analysis of masses in Cygnus X-1 based on the observed mass function $F = 0.245M_\odot$. (a) Mass of unseen companion M versus tilt angle i for various assumed masses M_c of the supergiant companion. (b) M versus M_c for various tilt angles i. Gray boxes indicate further observational constraints discussed in the text. The minimum possible mass for the unseen companion is given by the measured mass function $F = 0.245M_\odot$, which corresponds to the limit $M_\mathrm{c} \to 0$ and $i \to 90^\circ$.

Analysis of the observed velocity curve for the blue supergiant indicates $P = 5.6$ days and $K = 75\,\mathrm{km\,s^{-1}}$, which gives

$$F = \frac{PK^3}{2\pi G} = 0.245.$$

Solving the cubic equation and plotting M versus i and M versus M_c gives the graphs shown in Fig. 17.4.

- The spectrum–luminosity class of the blue supergiant is O9.7Iab, which permits its mass to be estimated from stellar systematics as 20–30 $M_\odot$.
- The tilt angle cannot be measured directly.
- However, detailed comparison with observed systematics for the system such as whether eclipses are seen permits it to be estimated as $i = 25 - 35^\circ$.



- These allow the acceptable ranges for a solution to be displayed as the gray boxes in the above figure.
- From this it may be concluded that the mass of the unseen companion lies in the range 10–20 $M_\odot$.
- No plausible equation of state supports a neutron star with $M > 2{-}3\,M_\odot$.

Hence it may be concluded that the unseen companion in Cygnus X-1 can only be a black hole.

A more comprehensive analysis concludes that

$$i = 27.1 \pm 0.8^\circ \qquad M_\mathrm{c} = 19.2 \pm 1,\, M_\odot$$

implying that

$$M = 14.8 \pm 1.0\, M_\odot.$$

- This more precise result is consistent with our simple estimate for Cyg X-1.
- It may also be noted that an extensive analysis has concluded that the black hole in Cyg X-1 has a spin greater than 95% of the maximal Kerr value.
- Thus, Cyg X-1 may be a *near-extremal Kerr black hole.*

Table 17.2: Black hole candidates in galactic X-ray binaries

X-ray source	Period (days)	$f(M)$	$M_{\rm c}(M_\odot)$	$M(M_\odot)$
Cygnus X-1	5.6	0.24	24–42	11–21
V404 Cygni	6.5	6.26	$\sim$0.6	10–15
GS 2000+25	0.35	4.97	$\sim$0.7	6–14
H 1705–250	0.52	4.86	0.3–0.6	6.4–6.9
GRO J1655–40	2.4	3.24	2.34	7.02
A 0620–00	0.32	3.18	0.2–0.7	5–10
GS 1124–T68	0.43	3.10	0.5–0.8	4.2–6.5
GRO J0422+32	0.21	1.21	$\sim$0.3	6–14
4U 1543–47	1.12	0.22	$\sim$2.5	2.7–7.5

A similar analysis has been carried out for many X-ray binaries in the galaxy.

- A summary of the cases that place the mass of the unseen companion well above the maximum mass for a neutron star or white dwarf is shown in Table 17.2.
- These binary systems are assumed to contain a black hole of mass M as the unseen companion.
- Even in the absence of further information on $M_{\rm c}$ and i, the measured value of the mass function defines the lowest possible mass for the unseen companion.
- For several entries in Table 17.2, $f(M)$ is well above the maximum mass thought to be possible for a neutron star or white dwarf.

17.7 Black Holes and Gravitational Waves

The first direct observation of gravitational waves (GW) in 2015 has opened a *new window on the Universe*.

- Gravitational waves are capable of probing *dark events* that might not be observable using the tools of traditional astronomy.
- Black holes are the quintessential dark objects, so GW astronomy is well suited to their study.
- Indeed the first two gravitational waves reported by the LIGO collaboration were each interpreted as resulting from the merger of binary black holes.

These gravitational-wave observations

- Provide the strongest evidence to date for the existence of black holes with masses comparable to stars.
- In addition, their detailed interpretation has begun to yield quantitative information about the black holes that were involved in the merger.
- This in turn establishes a new methodology to study late stellar evolution for massive stars.

Gravitational wave astronomy may be the most powerful method at our disposal for the study of black holes, as will be discussed more extensively in a later chapter.

17.8 Supermassive Black Holes

The center of the Milky Way lies in the constellation *Sagittarius (Sgr)*.

- The center coincides approximately with the radio source *Sgr A**.
- Sgr A* is weak by radio-galaxy standards but is the *strongest RF source in our galaxy*.
- The center of the galaxy is cloaked by dust but it can be *studied at IR wavelenths* that penetrate the dust.
- A number of stars have been tracked for more than two decades near the center of the galaxy.
- The most-studied is a 15 solar mass main sequence star denoted *S0-2* (also often called S2) that has been tracked since 1992.
- The star S0-2
  - is in a highly-elliptical *Keplerian orbit* with
  - Sgr A* *near a focus*.
- Thus Kepler's laws and the orbit of S0-2 may be used to deduce the mass of Sgr A*.

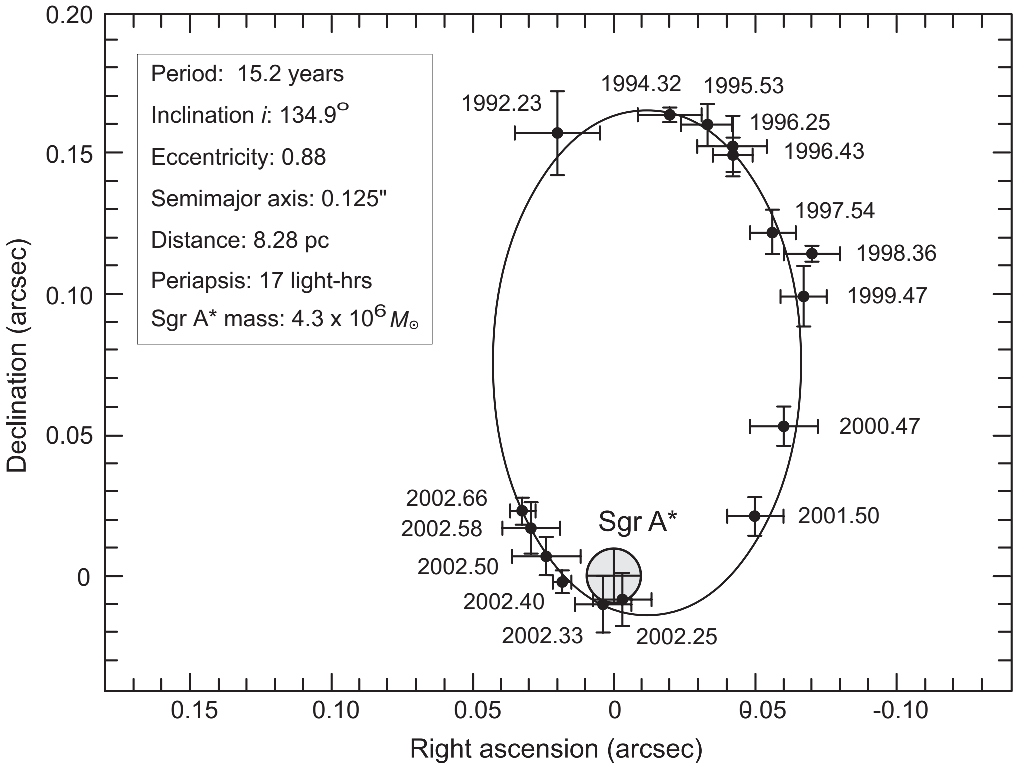
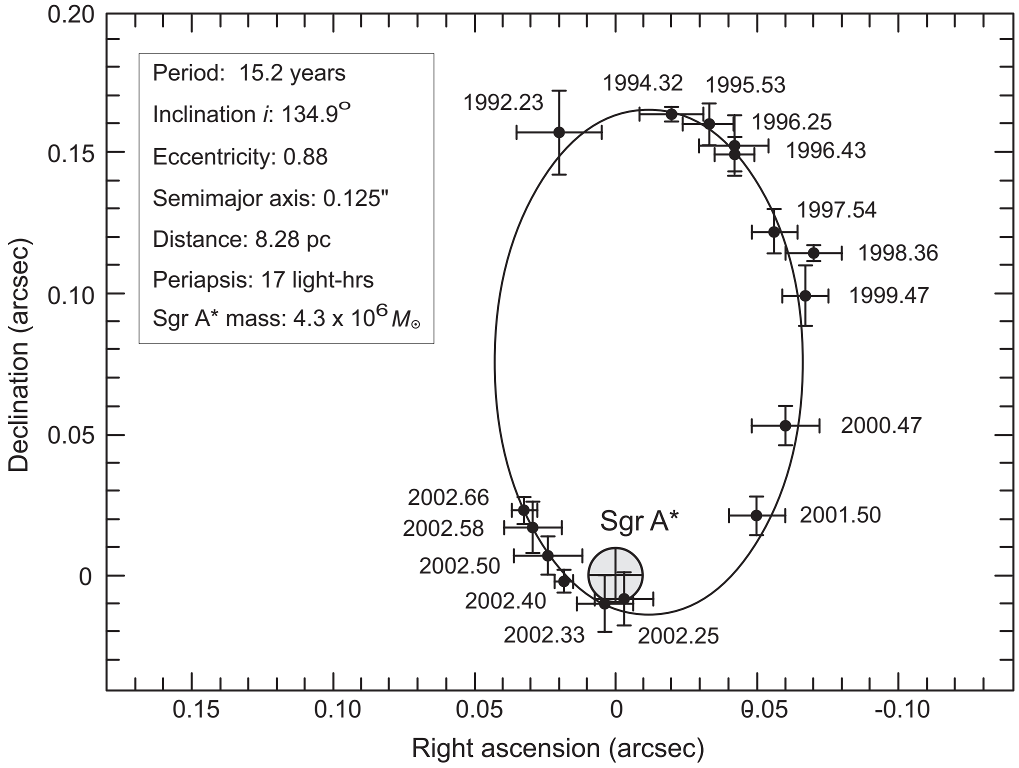


Figure 17.5: Orbit of S0-2 around Sgr A* through 2002. The filled circle indicates the position uncertainty for Sgr A* assuming a point mass located at the focus to be responsible for the orbital motion. The star completed this orbit in 2008 and the parameters displayed in the box are those obtained from the completed orbit. Periapsis is the general term for closest approach of an orbiting body to the center of mass about which it is orbiting.

Positions for S0-2 through 2002 are shown in Fig. 17.5.

- Dates are shown in fractions of a year from 1992.
- The orbit drawn in Fig. 17.5 corresponds to the *projection of the best-fit ellipse with Sgr A* at a focus*.
- At closest approach the separation of S0-2 from Sgr A* is only *17 light-hours*.



From fits to the orbit of S0-2 assuming Keplerian motion,

- The *mass inside the orbit* is $4.3 \times 10^6 M_\odot$.
- This mass is contained in a region that *cannot be much larger than Solar System* (and may be smaller).
- Within this region there is *little luminous mass*.

The simplest explanation is that the radio source Sgr A* *coincides with a* $4.3 \times 10^6 M_\odot$ *black hole* at the center of the Milky Way.

From extensive evidence based on observing *average star motion* near the centers of other galaxies

- The motion of the stars indicates *enormous amounts of invisible mass* at the centers of large galaxies.
- The simplest explanation is that *black holes containing millions to billions of solar masses* are common in the centers of galaxies.
- Whether such *supermassive black holes* form by
  - the *merger of many stellar black holes* created by stellar core collapse, or
  - by some process independent of stellar evolution like *direct collapse from gas clouds*

  is unknown at present.

Thus it is unclear whether supermassive black holes are directly relevant to our discussion of stellar evolution.

17.9 Intermediate-Mass and Hawking Black Holes

For completeness we remark briefly about two other possible classes of black holes that might not be connected very directly to issues in stellar evolution:

1. *Intermediate-mass black holes:* The evidence for intermediate-mass black holes (*hundreds to tens of thousands of solar masses*) has been inconclusive.
   - However in 2017 a pulsar was discovered orbiting an unseen mass concentration in the globular cluster 47 Tucanae.
   - The precise timing of the pulsar indicates that the magnitude of the unseen mass concentration is $2200^{+1500}_{-800}\,M_\odot$.
   - No electromagnetic signal has been detected, so if this is a black hole it must not be accreting.
   - Does this represent evidence for an intermediate-mass black hole?
2. *Hawking black holes:* Because of quantum effects,
   - Black holes can radiate their mass over time as *Hawking radiation.*
   - The emission rate is *negligible except for black holes of tiny mass* (say the mass of a proton).

   These are termed *Hawking black holes.*

The detailed properties and formation mechanisms for *intermediate-mass black holes* are not well understood.

- Hence it is not clear whether they have any connection to stellar evolution.
- For example,
  - do they form through clumping of stellar-mass black holes, and
  - is this an intermediate step in forming the supermassive black holes, or
  - do they collapse directly from gas clouds, independent of stellar evolution?

For *Hawking mini black holes*

- there is no observational evidence thus far, and
- if they exist they must have been formed in the incredibly high temperatures and densities of the big bang, *not in stellar processes*.

Thus Hawking black holes are of large potential interest for theories of quantum gravity, but they are not relevant for the present discussion.

17.10 Proof of the Pudding: Event Horizons

A compelling circumstantial case may be made for the existence of black holes.

- However, the black hole property that distinguishes it from anything else is its *event horizon,* and
- None of the evidence for black holes to date (2018) demands the existence of an event horizon.
- Therefore, irrefutable proof requires finding evidence for the event horizon of a black hole, which is obviously a considerable challenge.
- The best prospects are for the supermassive black hole at Sgr A*, which has a Schwarzschild radius of about $18 R_{\odot}$.
- Thus, seeing the event horizon of the Sgr A* black hole requires resolving an object of this size at a distance of about 8 kpc.

This may be possible soon, as very long baseline RF interferometry with arrays of broadly-dispersed radio telescopes can now achieve resolutions of this magnitude.

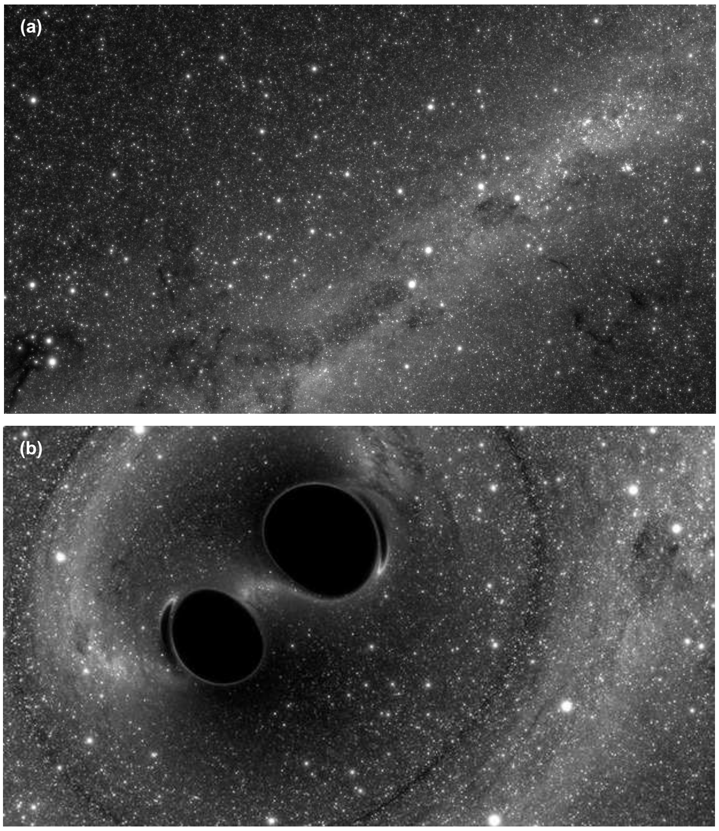
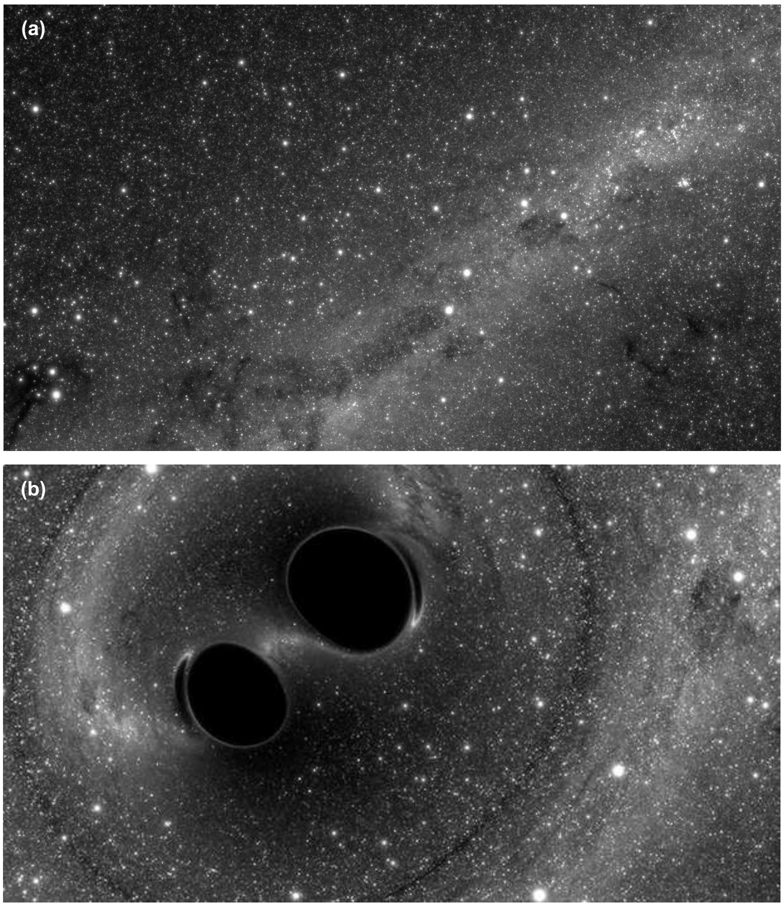


Figure 17.6: Computer simulation showing what the two black holes might have looked like just prior to merger in the gravitational wave event GW150914. (a) Background stars in the absence of the black holes. (b) Image including black holes. The ring around the black holes is an *Einstein ring,* which results from strong focusing by gravitational lensing of the light from stars behind the black holes. .

- A hint of what a resolved event horizon might look like comes from detailed analysis of data from the gravitational wave event GW150914 (corresponding to the merger of $29M_\odot$ and $36M_\odot$ black holes).
- A frame from a computer simulation of how the merger might have looked from nearby is shown in Fig. 17.6.



The jet-black shapes are the event horizons shadowing all light from behind.

- All stars are in the background but
- gravitational lensing in the strongly-curved space near the black holes severely distorts their apparent positions, and
- produces various lensed features around the event horizons and surrounding the black holes.

In the preceding image the black holes are assumed *isolated with no surrounding matter*.

- Hence the image is dominated by
  - the shadowing of the *black hole event horizons* and
  - strong *gravitational lensing effects*.
- In contrast, the black hole at Sgr A*
  - is in a dense cluster of stars and
  - is likely accreting surrounding matter and producing radiation from this accretion.
- The dominant observational feature may still be
  - the complete and sharply-defined shadowing of background light by the event horizon and
  - strong gravitational lensing near the horizon.
- However, the open question is how the environment of Sgr A* will distort this picture and
- whether the event horizon will still be identifiable in sufficiently-resolved observations.

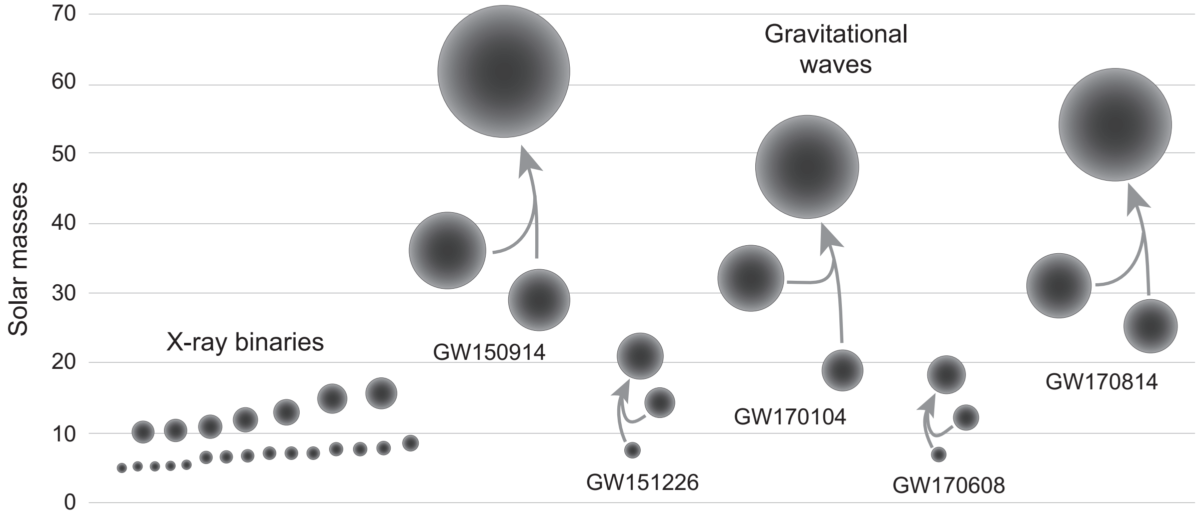


Figure 17.7: Summary of black hole masses determined from X-ray binary and gravitational wave (GW) data. Arrows indicate black hole mergers.

17.11 Summary of Measured Black Hole Masses

The most reliable methods for discovering stellar-size black holes and determining their masses are

- the mass-function analysis of *X-ray binaries* and
- Analysis of *gravitational waves* from black hole mergers.

Figure 17.7 summarizes masses for more than 35 black holes determined from these two types of analysis.

- These data constitute the strongest evidence now available for the existence of stellar-size black holes.
- It would be difficult to account for these data through any hypothesis other than that of black holes.

Part III

Accretion, Mergers, and Explosions

Chapter 18

Accreting Binary Systems

Observation suggests that most stars are in binary systems.

- When binary components are well separated they largely behave as isolated stars unless there are strong winds.
- However, if the semimajor axis of the orbit is small enough, mass may spill directly from one star onto the other. This is an example of *accretion*.
- Although accretion may not sound like a very exciting topic, in fact it is a critical ingredient in many of the most interesting phenomena in astrophysics.
- It plays this role either as
  - a mechanism initiating such phenomena (*novae* or *Type Ia supernovae*), or
  - as the primary power source (*supermassive black hole engines* that power quasars),

  or both (*high-mass X-ray binaries*).

We now investigate some accretion-driven phenomena.

18.1 Categories of Accretion in Binary Systems

It is useful to divide binary star accretion into two categories.

1. If the stars are sufficiently close together,
   - a gas particle "belonging" to one star may wander far enough from that star to be captured by the gravitational field of the other star.
   - This is termed *Roche-lobe overflow.*
2. Even if the two stars are not close enough together for Roche-lobe overflow to occur,
   - mass may be transferred between them if one star has a very strong wind blowing from its surface and
   - the second star captures particles from this wind.
   - This is termed *wind-driven accretion.*

As we shall see, these two methods of accretion tend to involve binary systems having very different total masses, with

- Roche-lobe overflow favored in *low-mass systems*.
- Wind-driven accretion favored in *high-mass systems*.

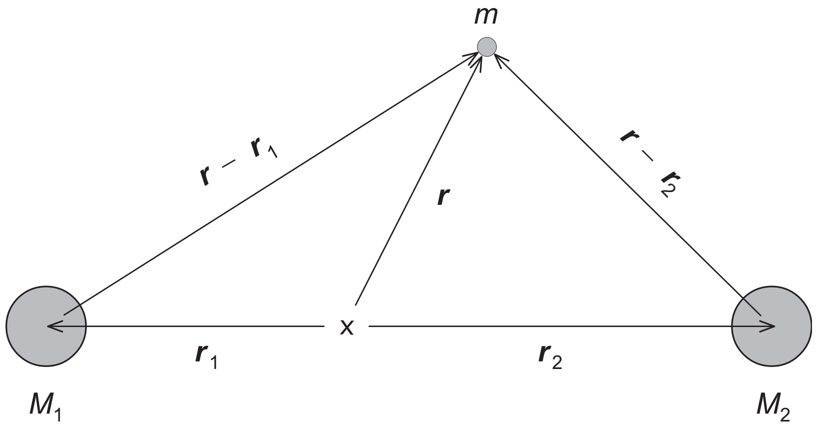


Figure 18.1: Three-body gravitational interaction. The restricted 3-body problem corresponds to assuming that m is much smaller than M_1 and M_2.

18.1.1 The Roche Potential

Consider the *restricted 3-body problem*, which is a 3-body gravitational problem where 2 of the masses may be considered to be much larger than a 3rd test mass (Figure 18.1).

- We are interested in the case where M_1 and M_2 are the two components for a binary star system in revolution around its center of mass and m is the mass of a gas particle.
- If we use a coordinate system rotating with the binary, the potential acting on the gas particle is termed the *Roche potential* $\Phi_{\rm R}(\boldsymbol{r})$, and is given by

$$\Phi_{\rm R}(\boldsymbol{r}) = -\frac{GM_1}{|\boldsymbol{r}-\boldsymbol{r}_1|} - \frac{GM_2}{|\boldsymbol{r}-\boldsymbol{r}_2|} - \tfrac{1}{2}(\boldsymbol{\omega}\times\boldsymbol{r})^2,$$

  where $\boldsymbol{\omega}$ is the frequency for revolution, and the other quantities are defined in Fig. 18.1.

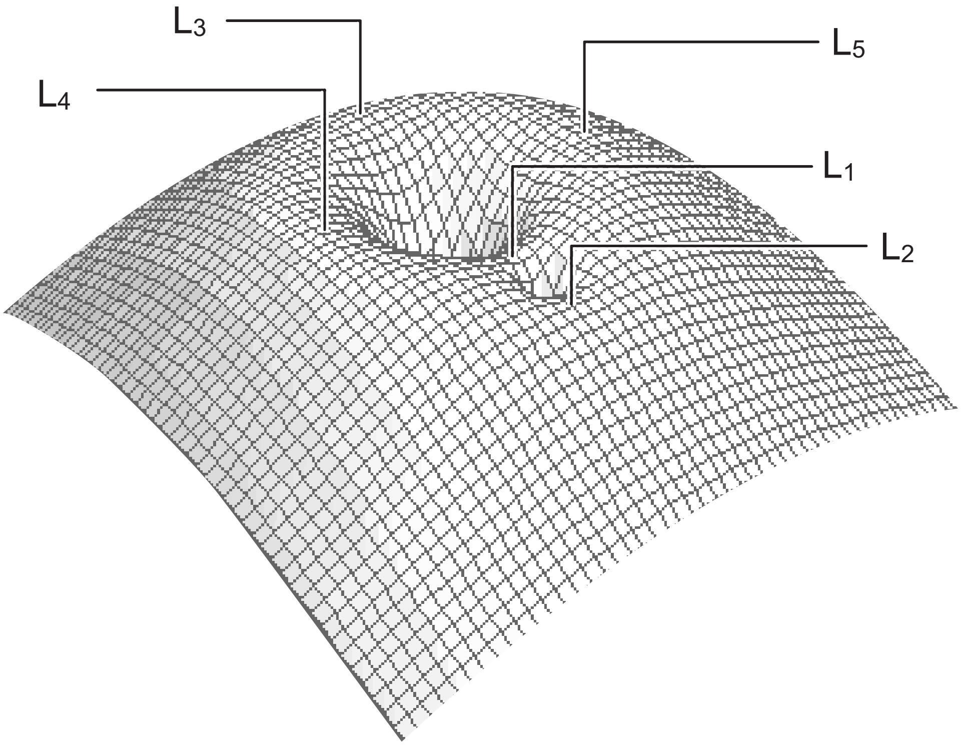


Figure 18.2: Gravitational potential and Lagrange points L_n for a binary system (surface courtesy of John Blondin).

A typical energy surface corresponding to the potential

$$\Phi_R(\boldsymbol{r}) = -\frac{GM_1}{|\boldsymbol{r}-\boldsymbol{r}_1|} - \frac{GM_2}{|\boldsymbol{r}-\boldsymbol{r}_2|} - \tfrac{1}{2}(\boldsymbol{\omega}\times\boldsymbol{r})^2,$$

is illustrated in Fig. 18.2.

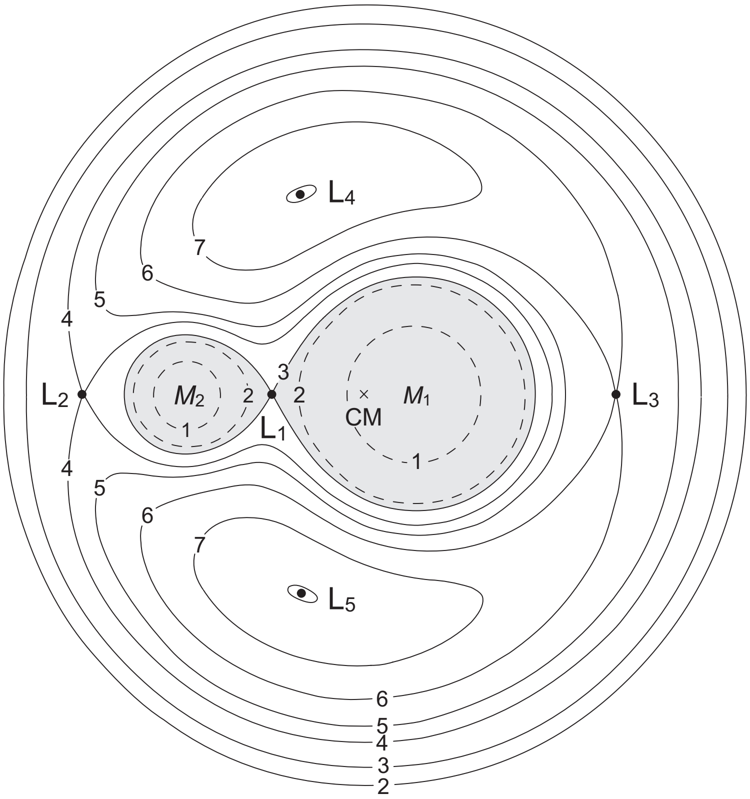


Figure 18.3: Gravitational potential contours for a binary system and the Lagrange points L_n. Dashed contours lie inside the Roche lobes (indicated in gray) and CM denotes the location of the center of mass.

A typical contour plot corresponding to the potential

$$\Phi_R(\boldsymbol{r}) = -\frac{GM_1}{|\boldsymbol{r}-\boldsymbol{r}_1|} - \frac{GM_2}{|\boldsymbol{r}-\boldsymbol{r}_2|} - \tfrac{1}{2}(\boldsymbol{\omega}\times\boldsymbol{r})^2,$$

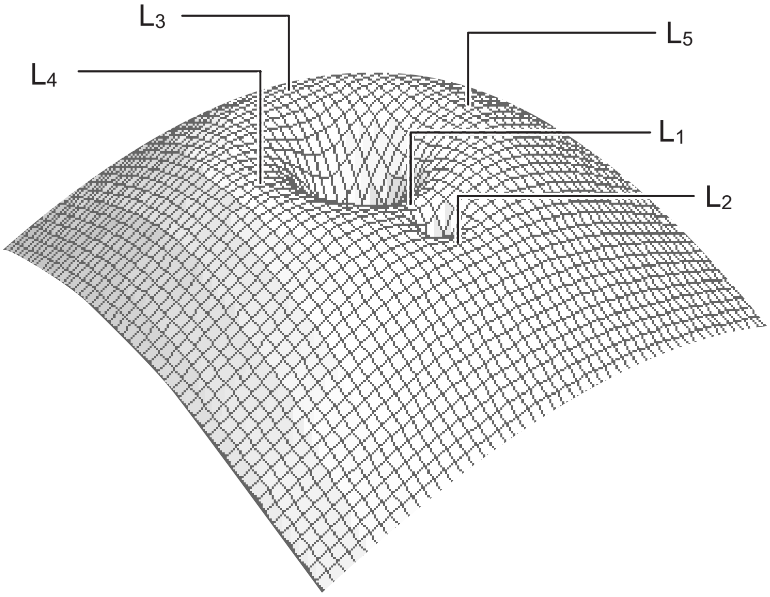
is shown in Fig. 18.3.

- The final term in

$$\Phi_{\mathrm{R}}(\boldsymbol{r}) = -\frac{GM_1}{|\boldsymbol{r}-\boldsymbol{r}_1|} - \frac{GM_2}{|\boldsymbol{r}-\boldsymbol{r}_2|} - \tfrac{1}{2}(\boldsymbol{\omega}\times\boldsymbol{r})^2,$$

  is required because *we have chosen a non-inertial coordinate system rotating with the binary*.

- It leads to *centrifugal and Coriolis (pseudo-) forces in the rotating frame*. For example, the fall-off of the potential



  at large distance is a consequence of *centrifugal effects*.

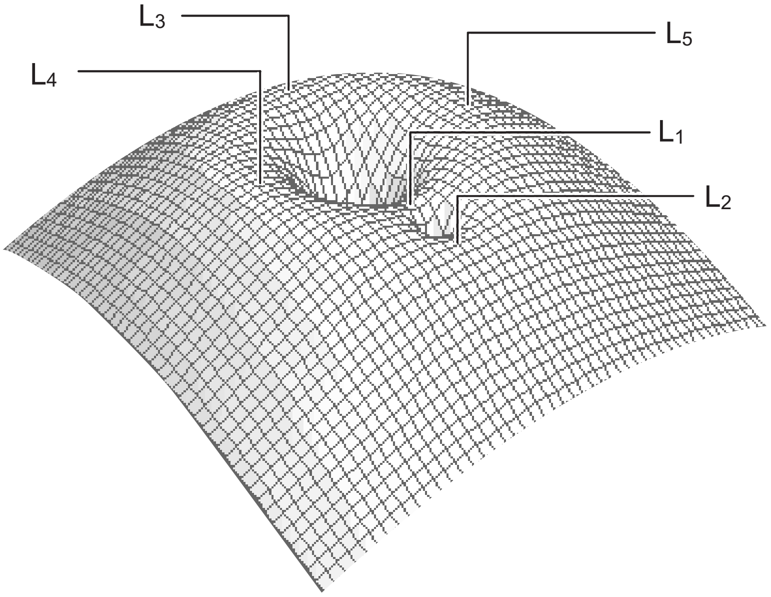
- The rotational frequency entering this equation is given by

$$\boldsymbol{\omega} = \sqrt{\frac{GM}{a^3}}\,\boldsymbol{e},$$

  where a is the semimajor axis, $M = M_1 + M_2$ is the total mass, and $\boldsymbol{e}$ is a unit vector normal to the orbital plane.

18.1.2 Lagrange Points

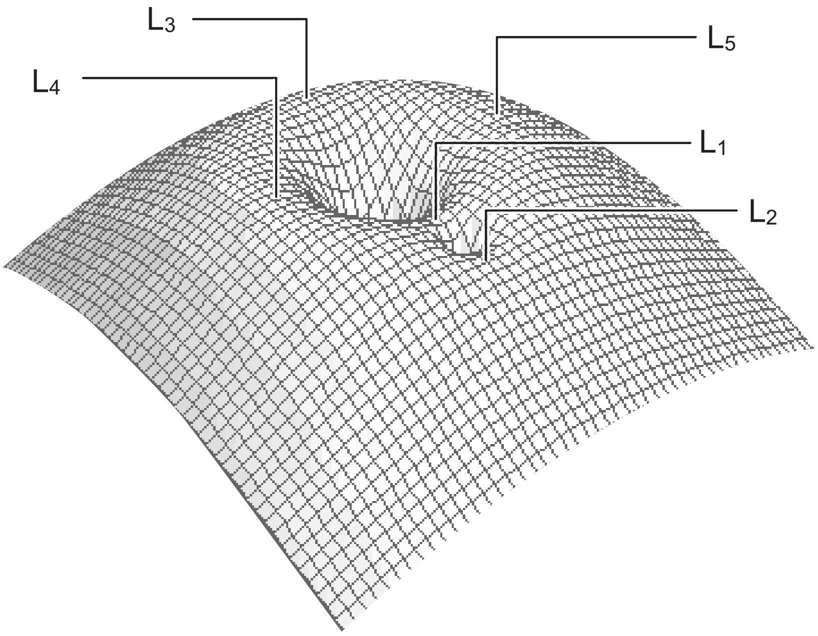
The five *Lagrange points* associated with the restricted 3-body problem are indicated in the following figure



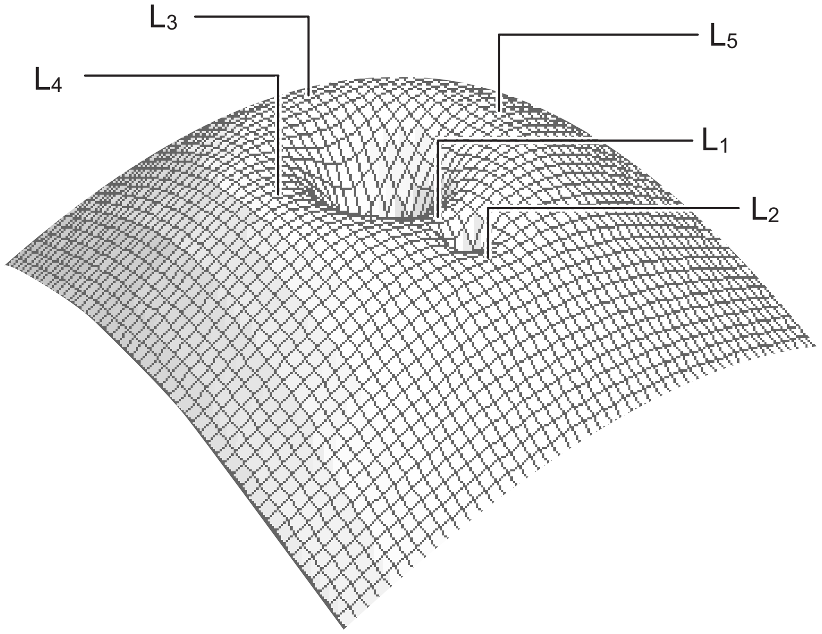
- These points correspond to the *five special points* in the vicinity of two large orbiting masses where a third body of negligible mass can orbit at a fixed distance from the larger masses.

- (Because at these points the gravity of the two large bodies is exactly balanced by the centripetal forces required for the small mass to rotate with them).

- The Lagrange points L_1, L_2, and L_3 lying on the line of centers for the two large masses are points of *unstable equilibrium* (saddle points of the potential).



- The points L$_4$ and L$_5$ are "hilltops" in the potential and seem to also be points of unstable equilibrium.
- However, for particular ranges of masses for the two large bodies, L$_4$ and L$_5$ are actually *stable* equilibrium points.
- The reason is the Coriolis force ofthe rotating frame.
- Basically, a particle rolling away from the hilltop at L$_4$ or L$_5$ experiences a Coriolis force that alters its direction.
- For favorable values of the parameters, the Coriolis deflection is sufficiently strong to put the particle into an orbit around the Lagrange point.



- For binary star systems the L_1 Lagrange point is of particular interest because mass flow between the stars can occur through the L_1 point.
- The L_2 point is also of potential interest, because mass overflow from star 1 to star 2 can in some cases overshoot and escape the system through the L_2 point.
- Such Lagrange points are also of considerable interest in the dynamics of natural and artificial objects in the Solar System, as discussed below.

Lagrange Points in the Solar System

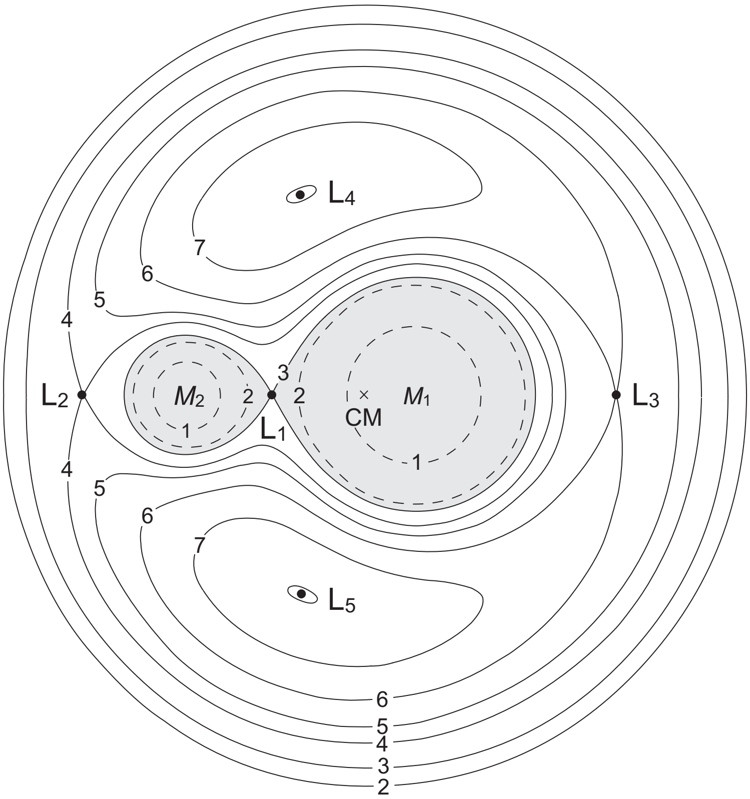
The Lagrange points play significant roles in the Solar System when one of the large masses is the Sun and one a planet.

- The Trojan Asteroids lie at the Jupiter–Sun L_4 and L_5 points, 60 degrees ahead and behind Jupiter in its orbit.
- The Solar and Heliospheric Observatory Satellite (SOHO) is parked at L_1 and the Wilkensen Microwave Anisotropy Probe (WMAP) at L_2 of the Earth–Sun system.
- The mythical "Planet X" of science fiction was purported to be at the L_3 point of the Earth–Sun system, and therefore always on the opposite side of the Sun from Earth.
- Dynamical analysis indicates that for the Earth-Sun system
  - the L_1 and L_2 points are unstable on a timescale of about 25 days;
  - thus the observatories parked there require small orbit corrections on that timescale to remain at the Lagrange points.
- The L_3 point for the Earth–Sun system is dynamically unstable on a 150-day timescale (bad news for Planet X!).
- The parameters of the Earth–Sun system, as for the Jupiter–Sun system, indicate that the L_4 and L_5 points are stable because of Coriolis forces.
- No Trojan-like asteroids have been found for Earth, but there is evidence for dust concentrations at L_4 and L_5.

18.1.3 Roche Lobes

One contour of the Roche potential intersects itself at the L_1 Lagrange point lying on the line connecting the center of mass for each star.

- The interior of this figure-8 contour defines a tear-drop shaped region for each star called a *Roche lobe*.
- The *Roche lobes* for the potential



are *shaded in gray*.

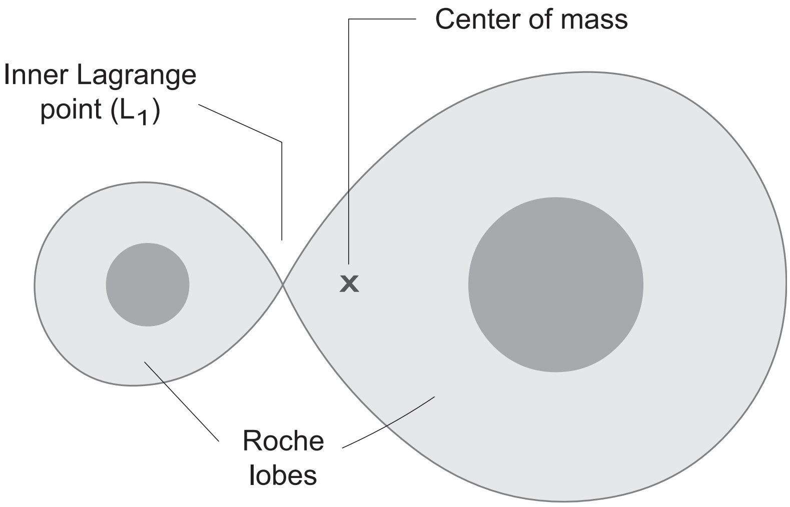


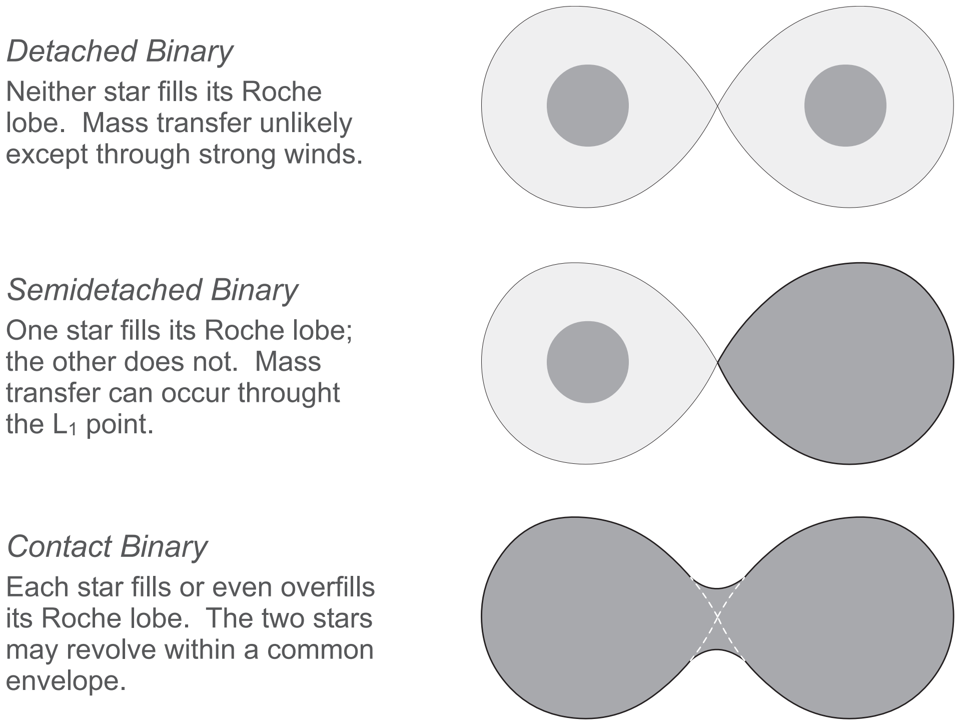
Figure 18.4: Roche lobes and the inner Lagrange point.

Fig. 18.4 illustrates Roche lobes more schematically.

- Roche lobes define the *gravitational domain* of each star.
- A gas particle within the Roche lobe of one star feels a stronger attraction from that star than from the other.
- It *"belongs" gravitationally* to the star unless there are instabilities (such as those responsible for winds) that upset the hydrostatic equilibrium.
- However, the L_1 Lagrange point is a saddle between the potential wells corresponding to the two stars.
- *A particle at L_1 belongs equally to both stars*, suggesting that mass transfer can be initiated if a star expands to fill its Roche lobe, thereby placing gas at the L_1 saddlepoint.

18.2 Classification of Binary Star Systems

The Roche lobes provide a convenient classification scheme for binary systems.



Detached Binary

Neither star fills its Roche lobe. Mass transfer unlikely except through strong winds.

Semidetached Binary

One star fills its Roche lobe; the other does not. Mass transfer can occur throught the L_1 point.

Contact Binary

Each star fills or even overfills its Roche lobe. The two stars may revolve within a common envelope.

Figure 18.5: Classification of binary systems.

1. In *detached binaries*, each star is within its Roche lobe.
2. In *semidetached binaries*, one star has filled its Roche lobe.
3. In *contact binaries* (W UMa stars), both stars have filled or overfilled their respective Roche lobes. This may
   - lead to a *"neck" between the stars*, or to
   - both stars orbiting within a *common envelope.*

During stellar evolution the classification of particular binary systems may change, as we discuss in the next section.

18.3 Accretion Streams and Accretion Disks

Let us now consider mass transfer through Roche-lobe overflow in a more quantitative manner. To do so requires a quantitative description of the gas flow between stars in a binary system.

18.3.1 Gas Motion

Gas motion is governed by the Euler equation

$$\rho\frac{\partial \boldsymbol{v}}{\partial t}+\rho\boldsymbol{v}\cdot\nabla\boldsymbol{v},=\nabla P+\boldsymbol{f}$$

$\boldsymbol{v}$ is velocity, ρ is density, P is pressure, and $\boldsymbol{f}$ is force density.

- This equation has the form

  $$(\text{mass density})\times(\text{acceleration})=(\text{force density}),$$

  and is a continuum version of Newton's second law.

- In a frame rotating with the binary at a frequency ω, the Euler equation takes the form

  $$\frac{\partial \boldsymbol{v}}{\partial t}+(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}=-\nabla\Phi_{\mathrm{R}}-2\boldsymbol{\omega}\times\boldsymbol{v}-\frac{1}{\rho}\nabla P.$$

- However, we shall now argue that many of the basic features of accretion through Roche-lobe overflow may be understood with only minimal calculation.

- These features follow largely from two observations:

  1. Mass transfer is extremely likely and highly efficient if a star fills its Roche lobe, and
  2. Usually, conservation of angular momentum for the transfered matter implies the formation of an *accretion disk* around the primary star.

18.3.2 Initial Accretion Velocity

Let us imagine that we view the accretion process from the vantage point of the compact primary onto which accretion takes place.

- In what follows we shall term the star onto which accretion takes place the primary and the other star the secondary of the binary.
- Notice that the primary then is not necessarily the brighter star (in most cases of interest it will be the less bright star).
- Tidal forces tend to quickly circularize orbits and to synchronize rotation with revolution in close binaries.
- Therefore we assume that the compact binary and the secondary star keep the same faces turned toward each other during the orbital period on a circular orbit.
- From our perch on the compact star the companion appears to be moving across the sky since it makes a complete circuit of the celestial sphere once each binary period.
- If the Roche lobe of the companion is filled so that matter comes across the L_1 point, it appears from our location on the compact star to have a large transverse component of motion because of the revolution of the binary system.

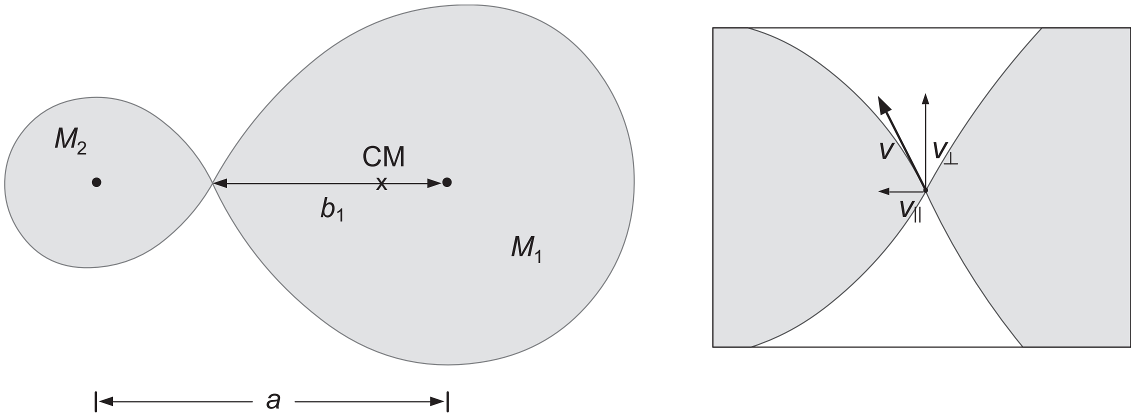
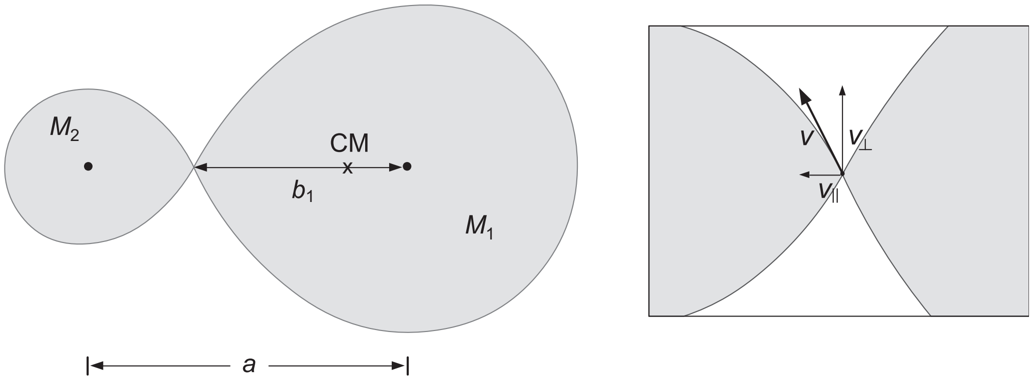


Figure 18.6: Angular momentum in Roche-lobe overflow.

- The relevant geometry is shown on the left side of Fig. 18.6.
- The components of velocity perpendicular to and parallel to the line of centers for the two stars are illustrated on the right side of Fig. 18.6.
- In the non-rotating frame, the perpendicular and parallel components of velocity for the stream of gas coming across the L_1 point satisfy

$$v_\perp \sim b_1 \omega \qquad v_\parallel \leq c_\mathrm{s}$$

where c_s is the local speed of sound in the vicinity of the L_1 point.



- The perpendicular component of the velocity may be estimated by using Kepler's third law. We may write

$$a = 2.9 \times 10^{11} m_1^{1/3} \left(1 + \frac{m_2}{m_1}\right)^{1/3} \left(\frac{P}{1\,\mathrm{day}}\right)^{2/3} \mathrm{cm},$$

  where $m_1 = M_1/M_\odot$ and $m_2 = M_2/M_\odot$.

- Taking $b_1 \sim \frac{1}{2}a$ and utilizing $\omega = 2\pi/P$, we find that

$$v_\perp \simeq 105 m_1^{1/3} \left(1 + \frac{m_2}{m_1}\right)^{1/3} \left(\frac{P}{1\,\mathrm{day}}\right)^{-1/3} \mathrm{km\,s^{-1}}$$

- The local sound speed may be approximated by

$$c_\mathrm{s} \simeq 10 \left(\frac{T}{10^4\,\mathrm{K}}\right)^{1/2} \mathrm{km\,s^{-1}}.$$

- For normal stellar envelopes $T \leq 10^5\,\mathrm{K}$ and therefore $v_\parallel \leq c_\mathrm{s} \leq 10\,\mathrm{km\,s^{-1}}$. Thus we obtain

$$v_\perp \sim \mathscr{O}\left(100\,\mathrm{km\,s^{-1}}\right) \qquad v_\parallel \sim c_\mathrm{s} \sim \mathscr{O}\left(10\,\mathrm{km\,s^{-1}}\right)$$

  for typical semidetached binaries having periods of days.

18.3.3 General Properties of Roche-Overflow Accretion

The preceding results have immediate implications for mass transfer through Roche-lobe overflow in binary star systems:

1. Since generally $v_\perp >> v_\parallel$, gas particles coming across the L_1 point will have *large angular momentum*.
2. Because $|\boldsymbol{v}| = (v_\parallel^2 + v_\perp^2)^{1/2} >> c_\mathrm{s}$, the accretion flow is in general *supersonic*.
   - Therefore, *pressure effects will be small*
   - (Supersonic flow has no time to react to pressure waves since they are limited to sound speed.)
   - Hence motion of the gas packets flowing across the L_1 point may be considered to be *ballistic*.
3. Because of the large angular momentum of particles the accretion stream will be *deflected by Coriolis effects*.
4. If it is deflected enough to miss the body of the primary, the accreting material will go into orbit around the primary, forming an *accretion disk*.
5. Because $v_\parallel \sim c_\mathrm{s} << v_\mathrm{ff}$, where v_ff is the velocity acquired by the particle as it is accelerated in the gravitational field of the primary, *initial conditions at the L_1 point will have little influence* on the accretion trajectory.
6. Thus the accretion stream should be *narrow* as it flows through the L_1 point into the Roche lobe of the primary.

With these assumptions, after passing through the L_1 point

- the test particle falls essentially freely in the gravitational potential of M_2, with the angular momentum that it had at the L_1 point.
- Thus, the test particle enters an approximately *elliptical orbit* in the plane defined by revolution of the binary.
- The test particles executing elliptical motion in the gravitational field of the primary form an *accretion disk* if
  - the transverse velocity of the particles entering the Roche lobe of the primary is sufficiently high that
  - the deflection of the accretion stream by the Coriolis effect causes it to miss the body of the primary.
- This is not always the case, but it typically will be in the most interesting situation where *the primary is a compact object*.

18.3.4 Disk Dynamics

The preceding discussion introduces most of the basic features of accretion by Roche-lobe overflow.

- However, the orbit of the test particle within the Roche lobe of the primary
  - is not actually a closed ellipse because of
  - gravitational perturbations,
  - most notably that caused by the presence of the secondary mass M_1.
- This deviation from a $1/r$ potential causes the ellipses to precess, leading to collisions of particles as orbits cross each other.

Collisions will have the following effects on the accretion disk:

- The collisions of particles will heat the gas in the disk, which can then emit energy as electromagnetic radiation.
- Shock waves presumably play a leading role in this heating because the velocities are generally supersonic.
- The accretion disk has limited opportunity to exchange angular momentum with external objects.
- Thus, the timescale for angular momentum transfer out of the disk is expected to be much longer than the timescale for radiating energy from the disk.
- As a consequence of the mismatch between timescales for radiating energy and transferring angular momentum, the particles in the disk will tend quickly to nearly circular orbits having the original angular momentum of the particle.
- (Circular orbits have the lowest energy for a given angular momentum.)
- The *circularization radius* R_{circ} is defined to be the orbit of lowest energy (that is, circular orbit) having the angular momentum of the test particles passing through L_1.
- It may be approximated by

$$\frac{R_{\text{circ}}}{a} = \frac{4\pi^2}{GM_1P^2}a^3\left(\frac{b_1}{a}\right)^4 = \left(1+\frac{M_2}{M_1}\right)\left(\frac{b_1}{a}\right)^4.$$

- Since the disk radiates energy,
  - some particles must descend lower into the gravitational potential of the primary to conserve energy.
  - To do so requires losing angular momentum.
  - But the timescale for transferring angular momentum from the disk is long compared with that for radiating energy,

  Thus the disk must transfer angular momentum internally:

  - Some particles in the disk must spiral inward while other particles spiral outward.
  - This net outward transfer of angular momentum implies that the disk is broadened both inward and outward around the circularization radius.

  A primary unresolved issue is the detailed mechanism by which an accretion disk accomplishes this internal redistribution of angular momentum.

The picture that emerges then is of a set of particles in the inner portion of the disk that slowly spiral inward on a series of nearly circular orbits of gradually decreasing radius in the binary plane.

> We may view an accretion disk as a device to slowly lower particles in the gravitational field of the primary until they accrete on its surface.

- The density and total mass of the accretion disk are typically low enough that we may safely ignore the self-gravity of the disk material.
- The particle orbits then tend to circular Kepler orbits with angular velocity

  $$\Omega(R) = \sqrt{\frac{GM_1}{R^3}},$$

  where M_1 is the mass of the primary and R is the radius of the orbit.
- For a Kepler orbit just grazing the surface of the primary at radius R_*, the binding energy of a gas packet of mass ΔM is

  $$E_{\text{bind}} = \frac{GM_1\Delta M}{2R_*}.$$

- In equilibrium the total luminosity of the disk must be

  $$L_{\text{disk}} = \frac{GM_1\dot{M}}{2R_*} = \frac{1}{2}L_{\text{acc}},$$

  where $\dot{M}$ is the accretion rate and L_{acc} is the accretion luminosity defined below.

Thus, half of the energy derived from accretion is radiated from the disk as the matter spirals inward toward the primary.

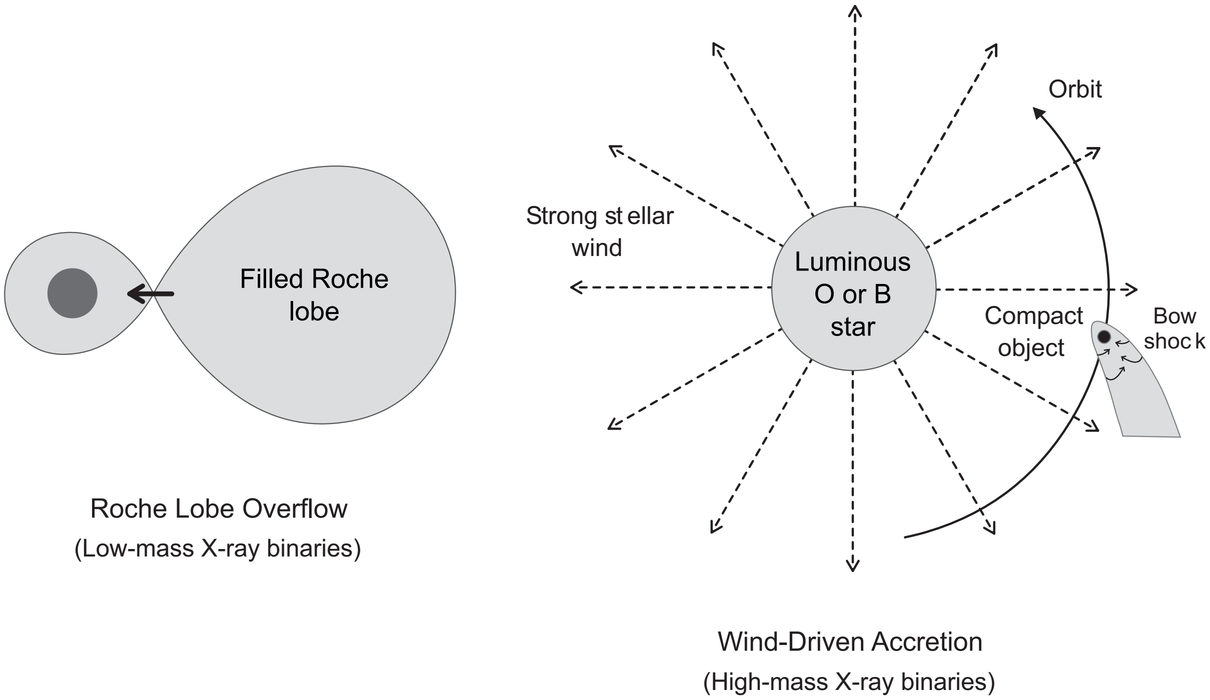


Figure 18.7: Roche-lobe overflow and wind-driven accretion. A bow shock is expected in the latter case because the wind flow is highly supersonic.

18.4 Wind-Driven Accretion

The schematic mechanisms for accretion driven by Roche-lobe overflow and by winds is illustrated in Fig. 18.7.

- Wind-driven accretion is far less well understood than is accretion by Roche-lobe overflow.
- Wind-driven accretion may be particularly important for those binary systems that contain an O or B spectral class star with a neutron star or black hole companion.
- These systems tend to be luminous sources of X-rays (see the discussion below of high-mass X-ray binaries).

The stellar wind from the early spectral class star is generally both *supersonic* and *intense*.

- Wind velocity may be approximated by *escape velocity*,

$$v_{\text{wind}} \simeq v_{\text{esc}} = \sqrt{\frac{2GM_*}{R_*}},$$

  where R_* is the radius and M_* the mass of the O or B star.

- This will typically be *several thousand kilometers per second*—far higher than the sound speed of $\sim$10 km s^{-1}.

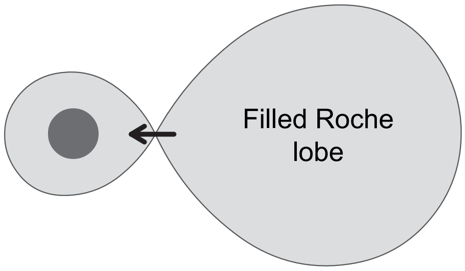
- The *rate of mass emission* from such hot, luminous stars is often as large as 10^{-6}–$10^{-5}\,M_\odot\,\text{yr}^{-1}$.

- The *highly supersonic particles* may be assumed to follow *ballistic trajectories*, which allows a simple estimate of the accretion rate on a compact companion.

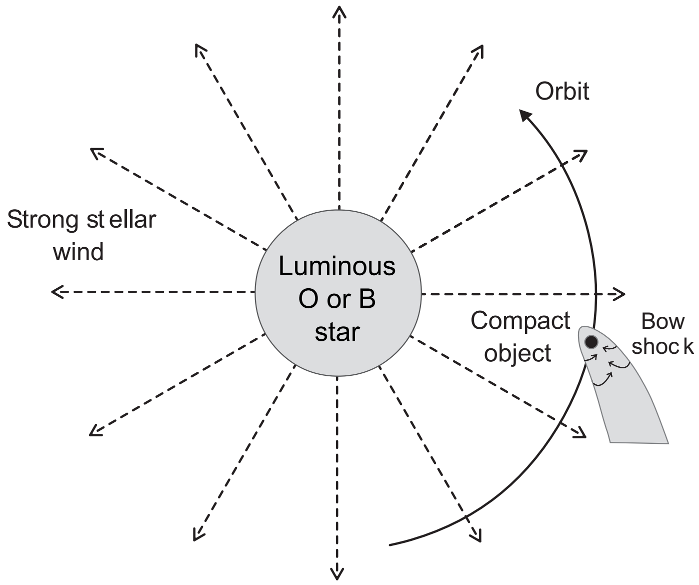
- Such estimates indicate that *wind-driven accretion is highly inefficient:* the accretion rate is 1000–10,000 times lower than the mass-loss rate from the companion.

- In contrast, *Roche-lobe overflow is highly efficient*, with close to 100% of the mass loss from one star accreting onto the other star in normal cases.

It is only the *high mass-loss rate* from the O or B star, and that *energy is emitted largely as X-rays* for neutron star or black hole companions, that permit wind-driven accretion to be observed.



Roche Lobe Overflow
(Low-mass X-ray binaries)



Wind-Driven Accretion
(High-mass X-ray binaries)

18.5 Classification of X-Ray Binaries

For binaries with highly compact remnants (neutron stars or black holes), persistent binary accretion seems to occur in only two general cases:

1. *High-Mass X-Ray Binaries* (HMXB) and
2. *Low-Mass X-ray Binaries* (LMXB).

These two extremes are summarized in the figure above and in the discussion below.

18.5.1 High-Mass X-Ray Binaries

Characteristics of high-mass X-ray binaries:

- Optical counterparts are typically *luminous O or B stars*.
- The optical luminosity from the system (dominated by visible and UV from the O or B star) is typically larger than the X-ray luminosity.
- They are commonly found in the galactic plane.
- They exhibit regular X-ray emission and transients with variation on timescales of minutes, but no large bursts.
- The X-ray spectrum is “hard”, with $kT \geq 15\,\mathrm{keV}$.
- HMXB are thought to consist of
  - a *neutron star* or *black hole*, and
  - a *high-mass* ($\geq 15M_\odot$) *companion* with a strong stellar wind, leading to *wind-driven accretion* on the compact object.
- Accretion rates $\dot{M} \sim 10^{-10} - 10^{-6}\,M_\odot\,\mathrm{yr}^{-1}$, with wind velocities $v \sim 2000\,\mathrm{km\,s}^{-1}$.

HMXB are often *very luminous X-ray sources* and were among the first X-ray binaries discovered in the galaxy.

The famous black hole candidate *Cygnus X-1* is an example of a *high-mass X-ray binary*.

18.5.2 Low-Mass X-Ray Binaries

In contrast to high-mass X-ray binaries, low-mass X-ray binaries exhibit the following characteristics:

1. LMXB have *faint blue optical counterparts* and the emission from the accretion disk may dominate over emission from the stars.
2. The optical luminosity is typically less than the X-ray luminosity by a factor of 10 or more, and the non-compact component is normally of spectral class A or later.
3. They are commonly parts of old stellar populations, spread out of the galactic plane and concentrated toward the galactic center.
4. They give rise to *strong X-ray outbursts*, with regular pulsations seen in only a few cases.
5. The X-ray spectrum is softer than for HMXB, with an effective $kT \leq 10\,\text{keV}$.
6. LMXB are thought to correspond to *binary systems having a compact star and a low-mass companion* ($\leq 2M_\odot$), with accretion onto the compact star by *Roche-lobe overflow*.

The *X-ray bursters* discussed later are examples of low-mass X-ray binaries.

18.5.3 Suppression of Accretion for Intermediate Masses

Thus, we see that

- HMXB correspond to *wind-driven accretion* from *high-mass companions* and
- LMXB correspond to *Roche-lobe overflow accretion* from *low-mass companions*,

with essentially *no X-ray binaries lying in between*.

This separation of mass scales can be understood as being caused by strong suppression of accretion onto compact objects from companions in the 2–15 $M_\odot$ range that arises for two reasons:

1. For companion masses lying in this intermediate range, *stellar winds from the companion are too weak* to drive significant X-ray luminosity from accretion.
2. For companion masses in this range, *Roche-lobe overflow accretion is quenched* because the mass transfer rates would become *super-Eddington*, effectively halting the accretion by virtue of the radiation pressure generated by the accretion.

Table 18.1: Energy released by accretion onto various objects

Accretion onto	Max energy released (erg g^{-1})	Ratio to fusion
Black hole	1.5×10^{20}	25
Neutron star	1.3×10^{20}	20
White dwarf	1.3×10^{17}	0.02
Normal star	1.9×10^{15}	10^{-4}

18.6 Accretion Power

The most spectacular consequence of accretion is that *it is an efficient mechanism for extracting gravitational energy*.

- The energy released by accretion is approximately

  $$\Delta E_{\text{acc}} = G\frac{Mm}{R},$$

  where M is the mass of the object, R is its radius, and m is the mass accreted.

- In Table 18.1 the amount of energy released per gram of hydrogen accreted onto the surface of various objects is summarized.

- From Table 18.1, we see that accretion onto very compact objects is a much more efficient source of energy than is hydrogen fusion.

- But accretion onto normal stars or even white dwarfs is much less efficient than converting the equivalent amount of mass to energy by fusion.

Let us assume for the moment, unrealistically, that

- all kinetic energy generated by conversion of gravitational energy in accretion is radiated from the system
- (we address the issue of efficiency for realistic accretion shortly).

Then the *accretion luminosity* is

$$L_{\text{acc}} = \frac{GM\dot{M}}{R} \simeq 1.3 \times 10^{21} \left(\frac{M/M_\odot}{R/\text{km}} \right) \left(\frac{\dot{M}}{\text{g s}^{-1}} \right) \text{erg s}^{-1},$$

if we assume a *steady accretion rate* $\dot{M}$.

Table 18.2: Some Eddington-limited accretion rates and temperatures

Compact object	Radius (km)	Max rate ($\mathrm{g\,s^{-1}}$)	T_{acc} (K)	kT_{acc} (eV)	Spectrum
White dwarf	~ 8000	8×10^{20}	$\sim 10^6$	~ 100	UV
Neutron star	~ 10	1×10^{18}	$\sim 10^7$	~ 1000	X-ray
$10 M_\odot$ black hole	~ 30	3×10^{18}	$\sim 10^7$	~ 1000	X-ray

18.6.1 Limits on Accretion Rates

The Eddington luminosity is

$$L_{\mathrm{edd}} = \frac{4\pi G M m_{\mathrm{p}} c}{\sigma},$$

with σ the effective cross section for photon scattering.

- For fully ionized hydrogen, we may approximate σ by the Thomson cross section to give

$$L_{\mathrm{edd}} \simeq 1.3 \times 10^{38} \left(\frac{M}{M_\odot} \right) \mathrm{erg\,s^{-1}}.$$

- If the Eddington luminosity is exceeded (in which case we say that the luminosity is *super-Eddington*), accretion will be blocked by the radiation pressure, implying that there is a maximum accretion rate on compact objects.
- Equating L_{acc} and L_{edd} gives

$$\dot{M}_{\mathrm{max}} \simeq 10^{17} \left(\frac{R}{\mathrm{km}} \right) \mathrm{g\,s^{-1}}$$

Eddington-limited accretion rates based on this formula are given in Table 18.6.2.

18.6.2 Accretion Temperatures

A crude estimate can be made of the accretion temperature for compact objects by

- assuming steady accretion at a rate $\dot{M}$ corresponding to the Eddington limit, and
- assuming that the accreted material equilibrates in a surface layer with a blackbody temperature T_{acc} given by

$$T_{\text{acc}} = \left(\frac{GM\dot{M}}{4\pi\sigma R^3} \right)^{1/4},$$

  where R is the radius and M the mass of the compact object, and σ is the Stefan–Boltzmann constant.

Accretion onto realistic compact objects is more complicated, involving

- accretion disks with possibly complex dynamics, and
- general relativistic effects that may not be negligible for accretion onto neutron stars and black holes.

Nevertheless, this simple estimate gives the right order of magnitude for accretion temperatures because they are determined primarily by the release of gravitational energy.

Some Eddington-limited accretion rates and temperatures

Compact object	Radius (km)	Max rate (g s^{-1})	T_{acc} (K)	kT_{acc} (eV)	Spectrum
White dwarf	~ 8000	8×10^{20}	$\sim 10^6$	~ 100	UV
Neutron star	~ 10	1×10^{18}	$\sim 10^7$	~ 1000	X-ray
$10 M_\odot$ black hole	~ 30	3×10^{18}	$\sim 10^7$	~ 1000	X-ray

The accretion temperatures and corresponding spectral regions for

- white dwarfs,
- neutrons stars, and
- a $10 M_\odot$ black hole

are also displayed in the table above. From this table we expect that

- Accretion on white dwarfs should lead to $T_{\mathrm{acc}} \sim 10^6$ K.
- Accretion on neutron stars and stellar-size black holes should lead to T_{acc} in excess of 10^7 K.

This corresponds to *spectra in the UV to X-ray region* for accretion on these objects.

18.6.3 Accretion Efficiencies

- For the gravitational energy released by accretion to be extracted,
  - it must be radiated as electromagnetic radiation or
  - matter must be ejected at high kinetic energy (for example, in AGN jets).
- Generally, we expect that such processes are *inefficient* and that only a fraction of the potential energy available from accretion can be extracted to do external work.
- This issue is particularly critical when black holes are the central accreting object, since
  - they have no "surface" onto which accretion may take place and
  - the event horizon makes energy extraction acutely problematic.

Accretion Efficiencies: From the previous equation for accretion power an *efficiency factor* η may be introduced through

$$L_{\text{acc}} = \frac{GM\dot{M}}{R} = \frac{GM}{Rc^2}\dot{M}c^2 = \eta\dot{M}c^2 \qquad \eta \equiv \frac{GM}{Rc^2},$$

where R is the effective accretion radius.

- Thus η is a measure of the *efficiency of converting mass to energy by accretion*.
- For accretion onto a white dwarf or neutron star we may take the radius of the object for R.
- For accretion on a spherical black hole we may assume that R is some multiple of the Schwarzschild (event horizon) radius, which is given by,

  $$r_{\text{s}} = \frac{2GM}{c^2} = 2.95\left(\frac{M}{M_\odot}\right)\text{ km},$$

  since any energy to be extracted from accretion must be emitted from outside that radius.
- Then for a spherical black hole

  $$L_{\text{acc}}^{\text{bh}} \equiv \frac{r_{\text{s}}}{2R}\dot{M}c^2 = \eta\dot{M}c^2 \qquad \eta = \frac{r_{\text{s}}}{2R}.$$

For a black hole a typical choice for R is the *radius of the innermost stable circular orbit* in the Schwarzschild spacetime, located at $3r_{\text{s}}$.

Efficiencies for Various Processes: Varying ranges of energy are available from processes that convert mass to energy.

- For *nuclear burning of hydrogen to helium* the mass to energy conversion efficiency is $\eta \sim 0.007$.
- For *accretion on compact spherical objects* like Schwarzschild black holes or neutron stars, reasonable estimates suggest $\eta \sim 0.1$.
- For *rotating, deformed (Kerr) black holes*, it is possible to be more efficient in energy extraction and efficiencies of $\eta \sim 0.3$ might be possible.

Example: The high energy-extraction efficiency provides a convincing argument that active galactic nuclei (AGN) and quasars must be powered by *rotating supermassive* ($M \sim 10^9 M_\odot$) *black holes*.

- For example, it is found that a quasar could be powered by accretion of as little as a few solar masses per year
- onto an object of mass $\sim 10^9 M_\odot$, and that
- this mass would occupy a volume the size of the Solar System or smaller.

Such properties are essential to explaining the energy sources of quasars and AGN.

18.6.4 Storing Energy in Accretion Disks

In addition to being a primary source of power for varied astrophysical phenomena, an accretion disk can function as a *storage reservoir for gravitational energy*.

- This can meter the energy release out over a much longer period than the dynamical timescale for direct collapse.
- For example, the long-period gamma-ray bursts to be discussed in later chapters *last as long as many tens of seconds* and are thought to be powered by the *collapse of the core of a massive star*.
- It is proposed that the core collapse leads to
  - a rotating black hole
  - surrounded by an accretion disk and
  - emitting ultrarelativistic jets on its rotation axis.
- The gamma-ray burst is then produced by the jets, energized partially by the accretion of matter from the disk.

Thus the accretion disk spreads part of the collapse energy out over tens to hundreds of seconds to power the long-period gamma-ray burst.

18.7 Some Important Accretion-Induced Phenomena

Accretion is a primary factor in a number of astrophysical phenomena, either as initiator, or as the primary power source, or as both. Let us summarize briefly some of these phenomena.

18.7.1 Cataclysmic Variables

Cataclysmic variables are *accreting binary systems in which accretion is onto a white dwarf*, with the accretion giving rise to a variety of outbursts depending on the circumstances.

- The most spectacular are *novae,* which correspond to a thermonuclear runaway under degenerate conditions that is triggered by the accumulation of a thin layer of accreted hydrogen on the surface of the white dwarf.
- This runaway ejects a rapidly expanding, hot shell, which is responsible for the sudden large increase in light output from the system.
- Accretion triggers the nova by dumping nuclear fuel on the surface of the white dwarf, but
- the nova outburst is powered by thermonuclear burning, primarily the hot-CNO cycle.
- Typical timescales for the thermonuclear runaway are 100–1000 seconds, but
- the increased light output by virtue of the expanding shell may last for decades as it expands and thins.
- Some novae have been observed to be recur over periods of years or decades.

18.7.2 X-Ray Bursters

X-ray bursts are events that occur in low-mass X-ray binaries. There are thought to be two categories.

- *Type II bursts* are less common and appear to be associated with fluctuations in the accretion rate, though convincing models are lacking.
- *Type I bursts* are more common and are characterized by X-ray luminosities that increase by factors of 10 or more over a period of a few seconds.
- Type I bursts exhibit large increases in X-ray luminosities during bursts but the integrated steady luminosity is larger than the burst luminosities by factors of 100 or more because bursts are of short duration.
- The accepted mechanism for Type I bursts is similar to that of a nova: thermonuclear runaway under degenerate conditions initiated by accretion, but the compact object onto which the accretion takes place is a neutron star rather than a white dwarf.
- The typical duration of the thermonuclear runaway in an X-ray burster is 1–10 seconds and bursts may recur on timescales of hours, days, or longer.
- The runaway is powered initially by hot-CNO reactions, but these break out into a series of rapid proton and α-particle capture interspersed by β^+ decays that produce proton-rich isotopes up to $\sim$ mass-100.

18.7.3 High-Mass and Low-Mass X-Ray Binaries

We have already discussed the distinction between low-mass and high-mass X-ray binaries. These systems, in which the compact object is either

- a neutron star or
- a black hole,

lead to a variety of X-ray emission associated with

- wind-driven accretion for the HMXB and
- Roche-lobe overflow for the LMXB

the X-ray bursters mentioned separately above are a particular example of low-mass X-ray binaries.

18.7.4 Type Ia Supernovae

Type Ia supernovae have been ascribed to two primary models, both likely involving accretion onto a white dwarf as the initiator of the explosion.

- *Single-degenerate scenario:* Accretion from a non-degenerate star onto a white dwarf triggers a runaway thermonuclear flash in the degenerate white dwarf matter that consumes the entire star.
- *Double degenerate scenario:* A binary white dwarf system spirals together and merges. Near merger one of the white dwarfs likely is tidally disrupted, forming a disk that accretes mass onto the other, triggering a thermonuclear runaway in degenerate white-dwarf matter.

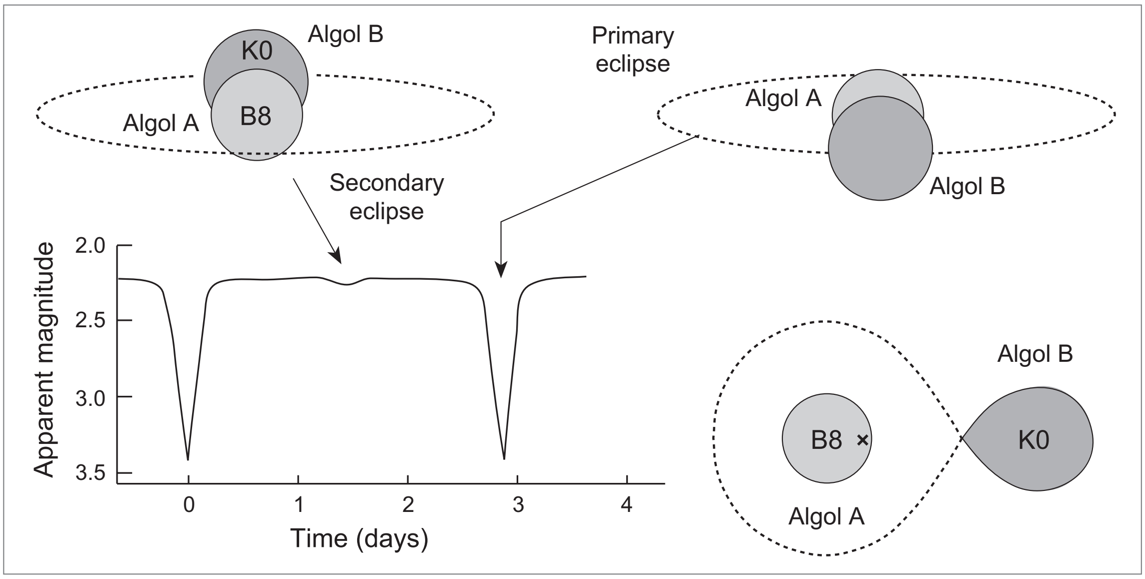
Type Ia supernovae will be discussed further in later chapters.

18.7.5 Supermassive Rotating Black Holes

Rotating black hole engines are thought to power active galaxies and quasars.

- These central engines produce far more power within a small region than can be accounted for easily by any source of energy other than gravitational.
- The standard paradigm is that these engines are powered by accretion onto supermassive, rotating black holes.

> On a stellar scale, gamma-ray bursts are believed to be powered in a similar way by accretion onto a rotating black hole produced either by
>
> - the core collapse of a massive star, or by
> - merger of two neutron stars (or a black hole and neutron star).
>
> This will be discussed further in later chapters.



18.8 Accretion and Evolution: the Algol Paradox

Mass is destiny for stars:

- Generally, the more massive a star is the faster it passes through all stages of it life.
- The evidence is overwhelming for this hypothesis but there are particular data sets that appear to contradict it.
- These apparent contradictions tend to involve stars interacting with another stars, either through accretion in a binary system or through collisions and mergers.

An interesting case is the Algol system shown above, where the more massive B8 star seems much less evolved than the less-massive K0 star.

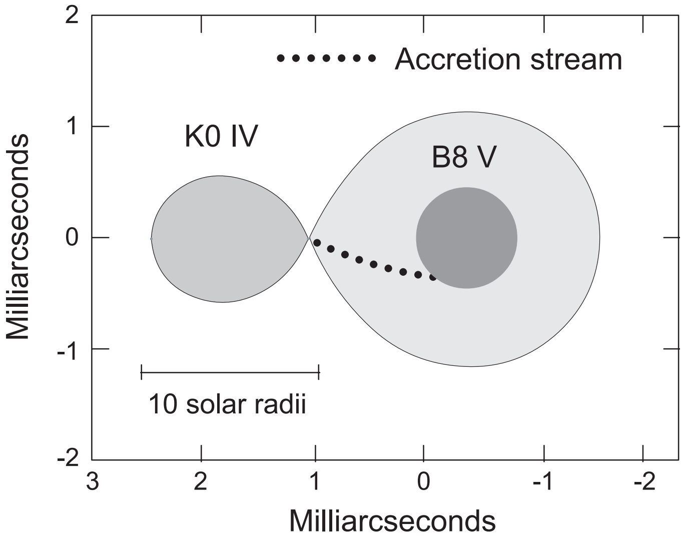


Figure 18.8: Mass transfer in the Algol system.

However, there is spectroscopic evidence that

- weak accretion is occuring in the Algol system
- (and that the accretion is directly into the body of the primary rather than through an accretion disk).

Therefore, we may ask whether it is possible that accretion in the Algol system is distorting our picture of which star is the older star.

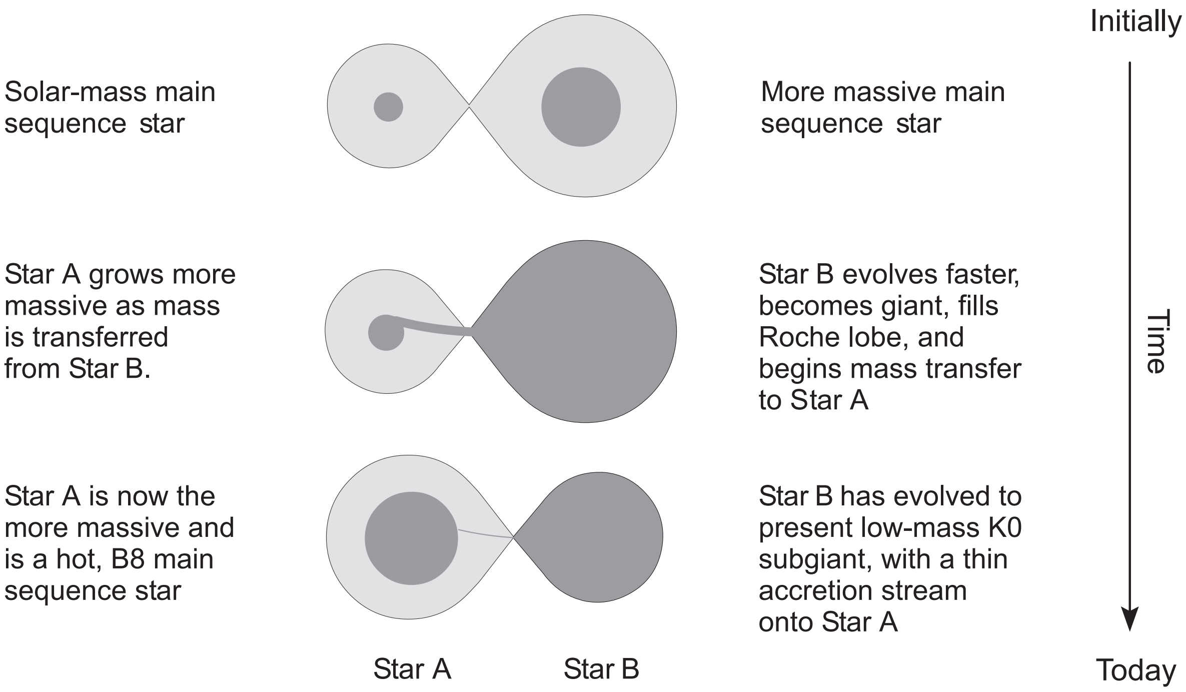
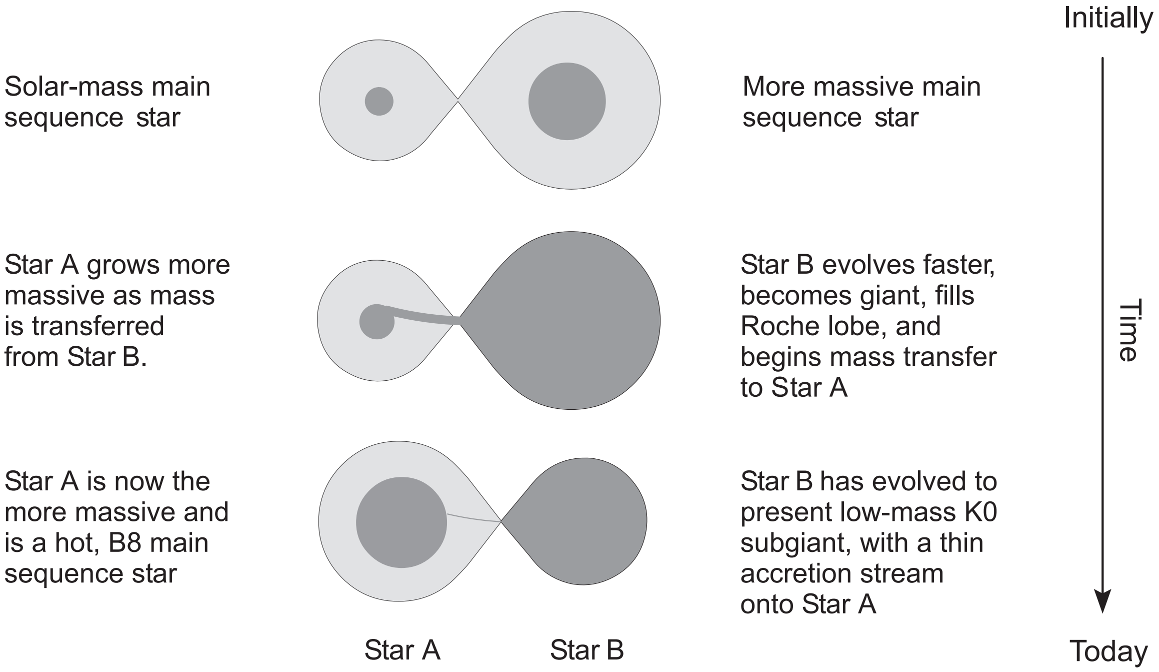

Initially
Solar-mass main sequence star
More massive main sequence star
Star A grows more massive as mass is transferred from Star B.
Star B evolves faster, becomes giant, fills Roche lobe, and begins mass transfer to Star A
Time
Star A is now the more massive and is a hot, B8 main sequence star
Star B has evolved to present low-mass K0 subgiant, with a thin accretion stream onto Star A
Star A
Star B
Today

Figure 18.9: Resolution of the “Algol paradox”.

This *Algol paradox* is thought to be resolved by the evolutionary sequence depicted in Fig. 18.9, which indicates that previous mass transfer has altered the system from what would have been expected for the evolution of isolated stars.


Initially
Solar-mass main sequence star
More massive main sequence star
Star A grows more massive as mass is transferred from Star B.
Star B evolves faster, becomes giant, fills Roche lobe, and begins mass transfer to Star A
Time
Star A is now the more massive and is a hot, B8 main sequence star
Star B has evolved to present low-mass K0 subgiant, with a thin accretion stream onto Star A
Star A
Star B
Today

- Initially the present K0 star (star B in figure) was a more massive main sequence star that evolved faster than its less massive companion (present B8 star, denoted star A).

- This initially more massive star evolved off the main sequence, expanded to fill its Roche lobe, and began rapid mass transfer to its companion.

- Over time enough mass transfer occurred to make the companion more massive and the present K0 star less massive, and accretion has diminished to a trickle.

- In some binary systems there is evidence for such mass transfer occurring first in one direction and then in the other, modifying substantially the evolution of both stars.

We may expect that the B8 star will fill its Roche lobe and begin mass transfer back to the K0 star at some point in the future.

A related issue may be *"blue stragglers"* in clusters, where

- main sequence stars more blue (earlier spectral class) than the turnoff point for the cluster are observed.
- According to standard evolutionary models, such stars *should already have evolved off the main sequence*.
- However, the presence of blue stragglers could be explained if they are not isolated stars but
- instead are stars that have interacted strongly with another star,
  - either with a companion from a binary system, or
  - by collision encounter in the dense environment of the cluster.

Then our normal evolutionary picture would be skewed because *these stars have not always been as massive as they presently are*.

Chapter 19

Nova Explosions and X-Ray Bursts

- Some stars appear to increase their brightness suddenly at visible (and other) wavelengths by large amounts over a period of days.
- Then they dim slowly back to obscurity over a matter of months.
- The increase in brightness can be as large as factors of a million.
- We call such a star a *nova.*
- Furthermore, many star systems are observed to be strong sources of X-rays.
- In some cases these X-ray sources are relatively steady.
- In others the emission of X-rays can come in sudden *X-ray bursts* superposed on a background emission

What is the nature of these nova and X-ray burst events, and what are the energy sources that power them?

- Strong clues are provided by the observation that both of these kinds of events seem to be associated with binary star systems.
- In this chapter we shall discuss novae and X-ray bursts in more detail and argue that they are caused by a similar mechanism:

> Novae and X-ray bursts are caused by a *thermonuclear runaway* triggered by *accretion onto a compact object.*

- The primary difference is that the compact object is
  - a *white dwarf* in the case of the *nova* and
  - a *neutron star* in the case of the *X-ray burst*.

19.1 Novae: Periodic Outbursts in Binary Systems

The nature of a nova event is suggested by three key observations.

1. Novae seem to be associated with *binary systems* in which one star is a *white dwarf*.
2. Doppler shifts indicate an *expanding shell of gas* emitting the light being observed from a nova.
3. There are *recurrent novae* (novae that repeat after some period of time).

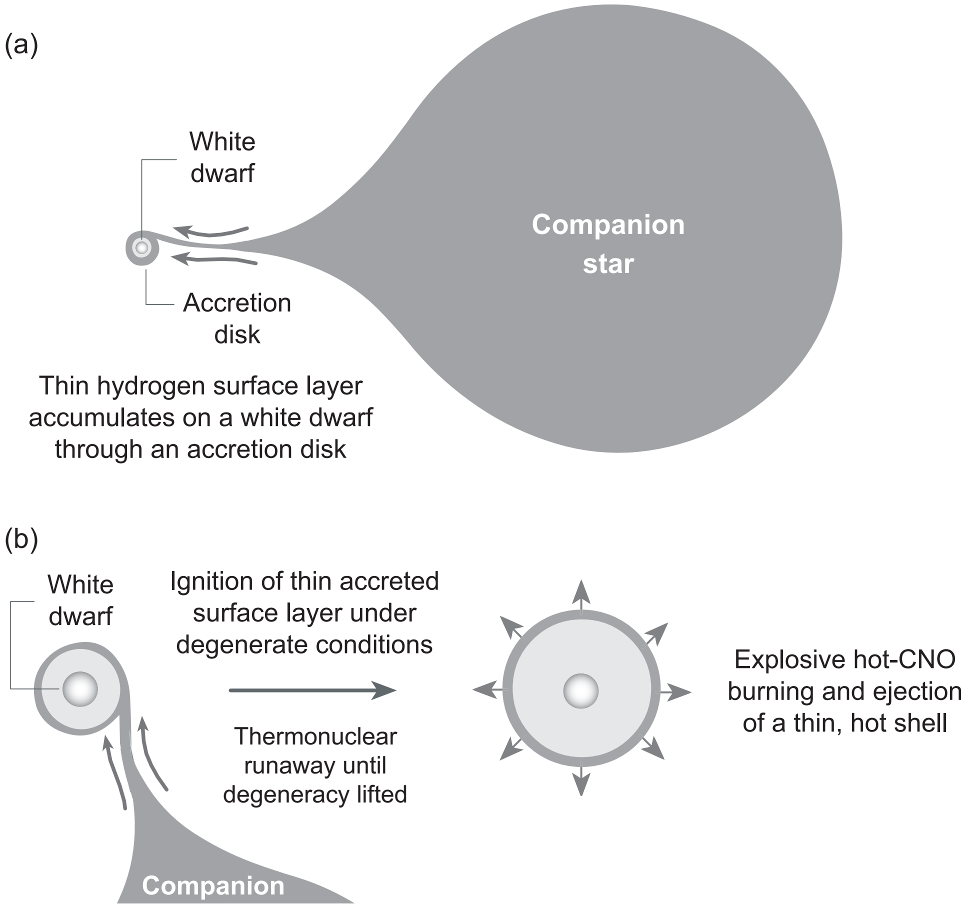
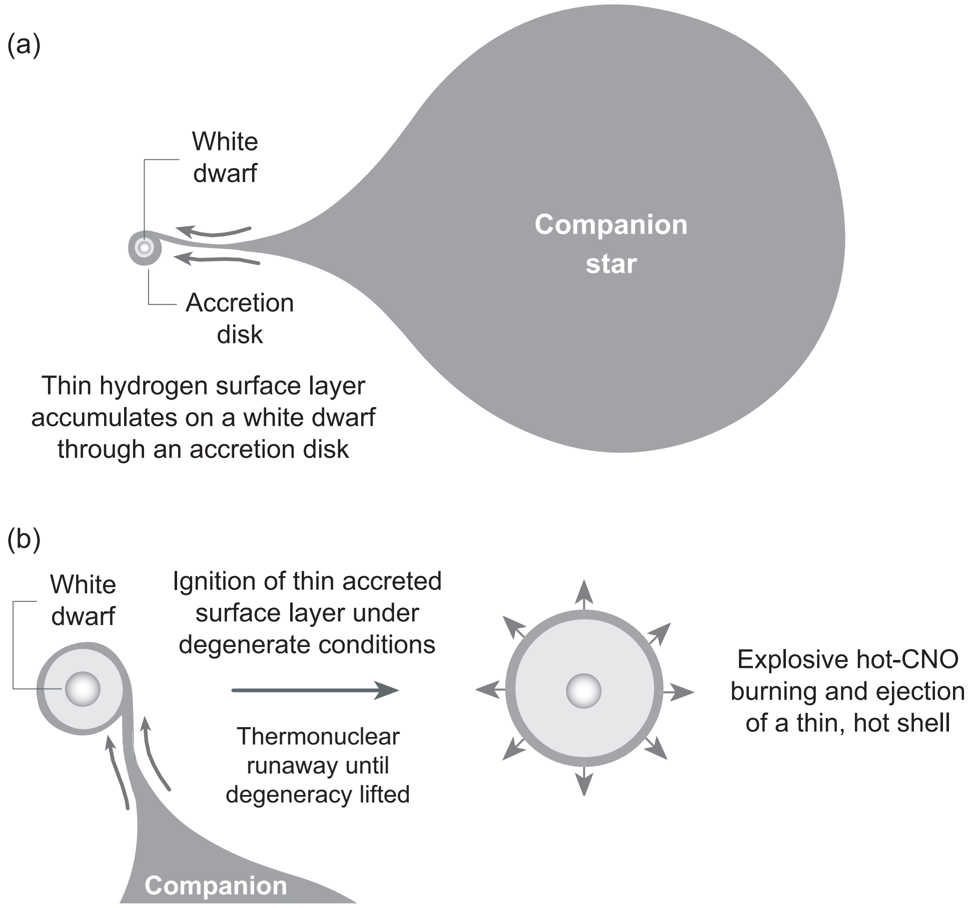

(a)
White dwarf
Accretion disk
Companion star
Thin hydrogen surface layer accumulates on a white dwarf through an accretion disk
(b)
White dwarf
Ignition of thin accreted surface layer under degenerate conditions
Thermonuclear runaway until degeneracy lifted
Explosive hot-CNO burning and ejection of a thin, hot shell
Companion

Figure 19.1: (a) Binary accretion leading to a nova outburst. (b) Hydrogen accumulated on the surface of the white dwarf can ignite in a thermonuclear runaway, blowing off a thin shell and producing the nova outburst.

This suggests the nova mechanism illustrated in Fig. 19.1,

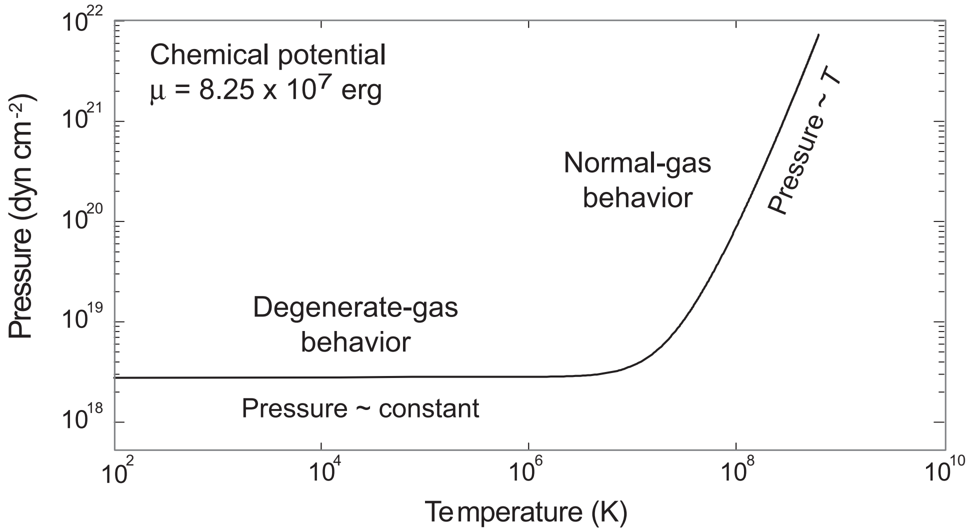
- A nova can occur in a binary system in which one star is noncompact, with a white dwarf companion accreting from the noncompact star.
- Matter from the first star accretes in a thin layer on the surface of the white dwarf.


(a)
White dwarf
Accretion disk
Companion star
Thin hydrogen surface layer accumulates on a white dwarf through an accretion disk
(b)
White dwarf
Ignition of thin accreted surface layer under degenerate conditions
Thermonuclear runaway until degeneracy lifted
Explosive hot-CNO burning and ejection of a thin, hot shell
Companion

- Eventually this layer ignites in a *thermonuclear explosion* under *degenerate conditions*.
- The resulting *thermonuclear runaway* (recall the earlier discussion of the helium flash in red giant stars and see the following box) blows a thin surface layer off into space.
- This causes a *large rise in light output* from the system.

Degeneracy and Thermonuclear Runaways in Novae

The equation of state for a gas of electrons under conditions expected in novae is illustrated in the following figure



- At low T the equation of state is degenerate and the pressure is essentially independent of the temperature.
- At high T the degeneracy is lifted and the pressure increases with temperature, as expected for an ideal gas.
- Thus, a thermonuclear reaction ignited in degenerate matter on the surface of the white dwarf becomes a runaway.
- This continues until the temperature rises sufficiently to break the degeneracy and produce a pressure increasing rapidly with temperature.
- This blows off the hot burning surface layer in a rapidly expanding thin shell.

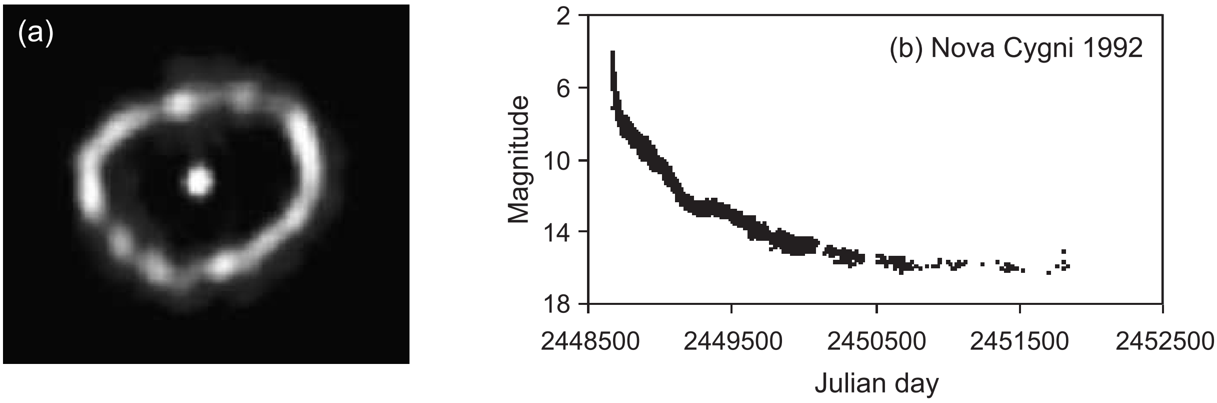


Figure 19.2: The expanding shell of gas around Nova Cygni 1992, two years after the nova explosion.

19.1.1 Nova Cygni 1992

Figure 19.2 shows the shell ejected by Nova Cygni 1992, as imaged by the Hubble Space Telescope two years after the explosion was first observed.

It is common to name novae using the word “Nova", followed by the constellation and the year the outburst was first observed. This nova was observed in the constellation Cygnus in 1992.

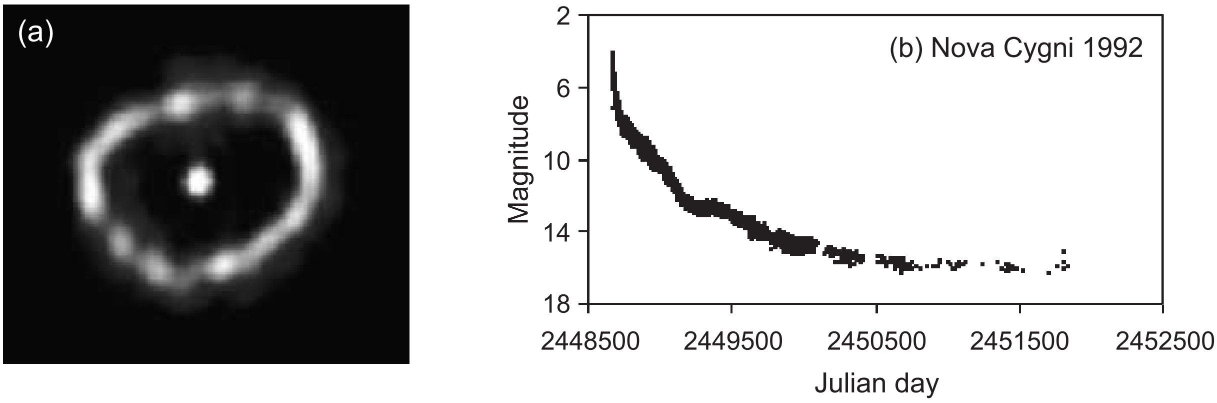


Figure 19.3: Visual lightcurve of Nova Cygni 1992 from the AAVSO International Database. The Julian dates on the bottom axis span August 31, 1991 to August 13, 2002.

The lightcurve (brightness versus time) for Nova Cygni 1992 is shown in Fig. 19.3(b).

- Nova Cygni 1992 was the brightest nova observed in recent years, and
- was visible without a telescope at its peak.

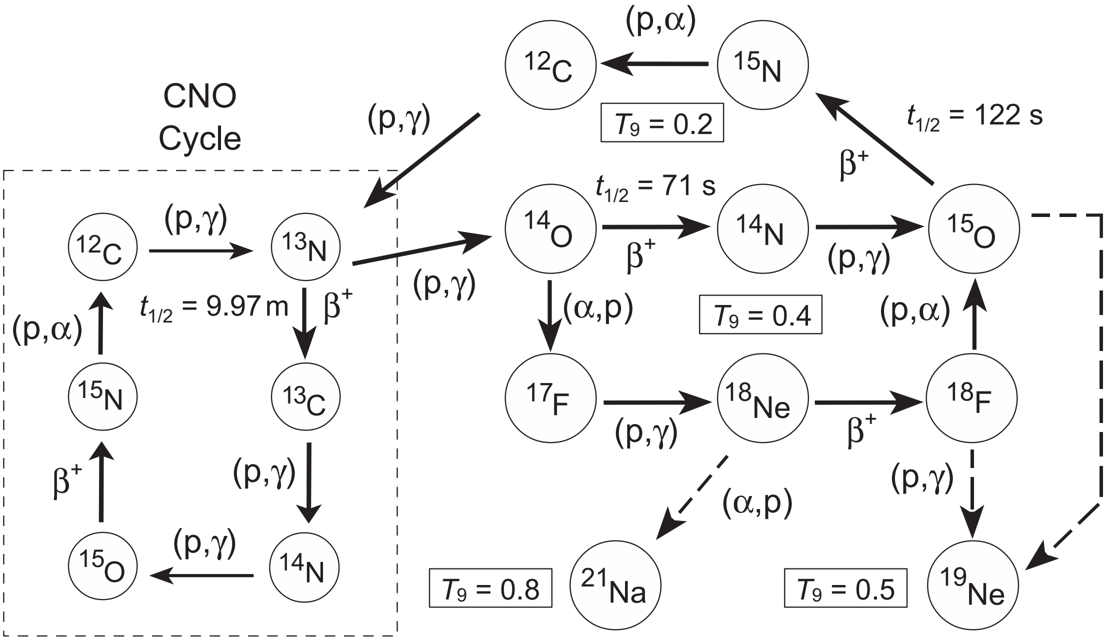


Figure 19.4: The hot CNO cycle and its relationship to the CNO cycle.

19.1.2 The Hot CNO Cycle

Given that a nova corresponds to a thermonuclear runaway on the surface of a white dwarf under degenerate conditions, what is the actual sequence of nuclear reactions that is responsible for the explosion?

- A nova explosion is powered by an extension of the CNO cycle at higher temperatures to a wider set of reactions called the *hot CNO cycle.*
- Figure 19.4 illustrates the relationship between the CNO and hot-CNO cycles.

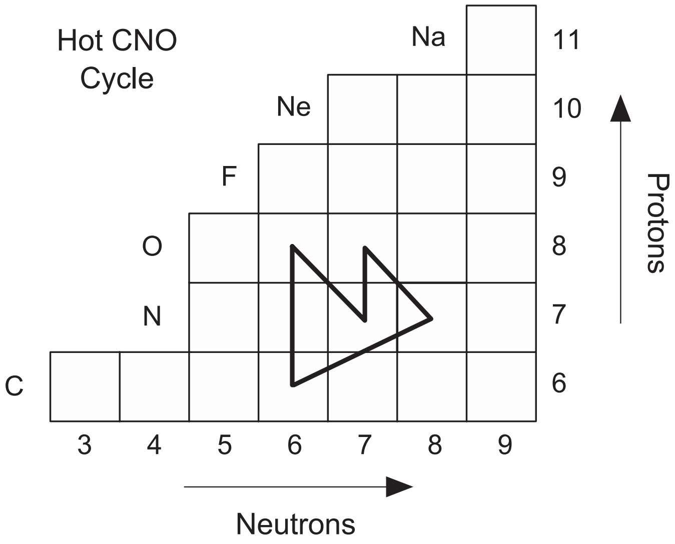


Figure 19.5: Main branch of the hot CNO cycle illustrated as a closed path in the proton–neutron plane. See also Fig. 19.4.

Alternatively, we may represent the hot CNO cycle in isotope space as in Fig. 19.5.

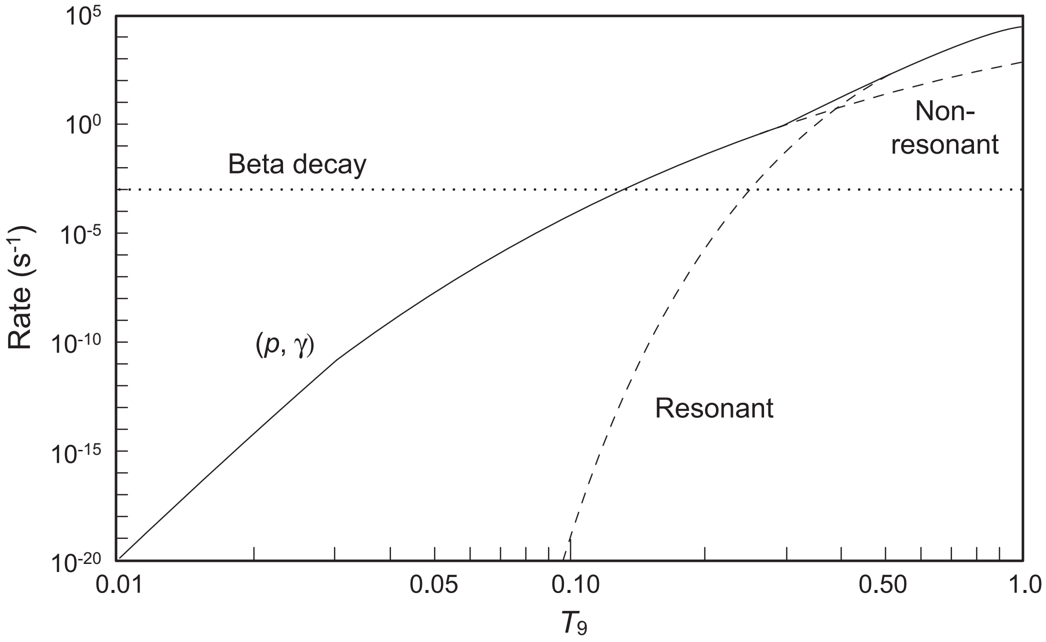
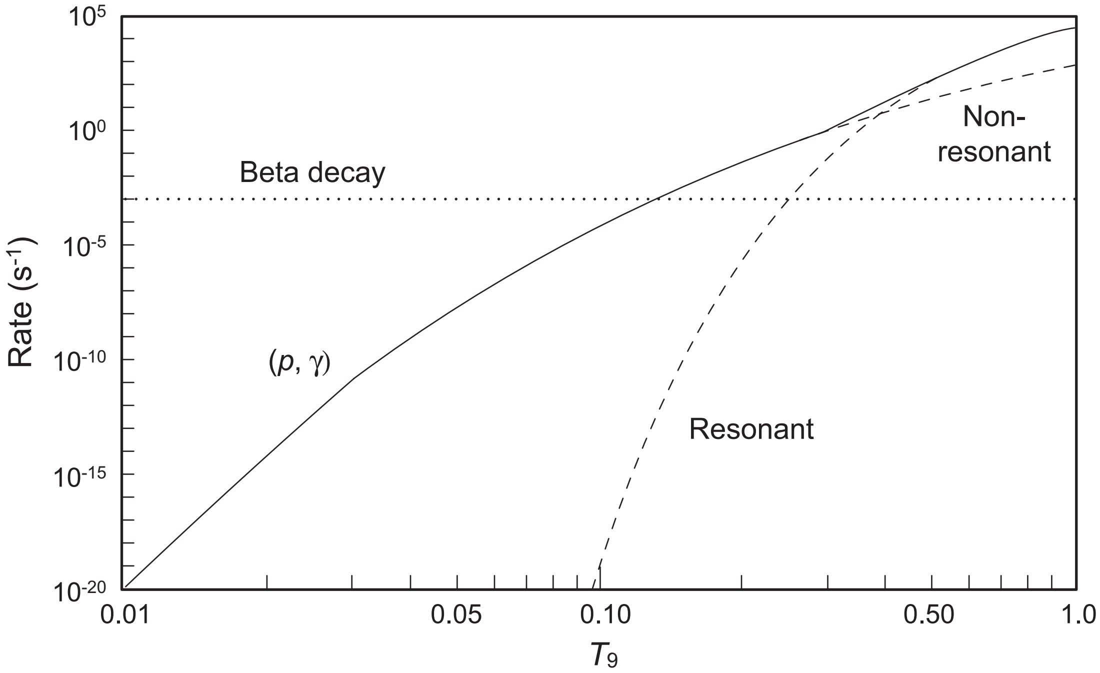


Figure 19.6: Competition of proton capture and β-decay in breakout from the CNO to hot CNO cycle. The solid line represents to total rate and dashed lines indicate resonant and non-resonant contributions to the total.

The transition from the CNO cycle to the hot CNO cycle is initiated by a proton capture reaction on ^{13}N.

- Whether the hot CNO cycle is populated is a strong function of temperature and its influence on the competition between proton capture and β-decay, as in Fig. 19.6.
- The β-decay of ^{13}N, which keeps the reaction flow within the CNO cycle, is essentially *independent of temperature*.
- However, ^{13}N can also capture a proton to make ^{14}O, which initiates the breakout into the hot CNO cycle.
- This reaction has an *extremely strong temperature dependence*, since it is inhibited by a Coulomb barrier.



- At low temperatures the β-decay wins.
- But for temperatures in excess of $T_9 \sim 0.1$ the proton capture reaction begins to compete strongly and quickly dominates with even small increases in temperature.
- The rising temperature of the initial nova outburst triggers this breakout into the hot CNO cycle and the nova is powered by the corresponding energy that is released.

Nuclear burning through the hot CNO cycle is often termed *explosive hydrogen burning*.

19.1.3 Recurrence of Novae

The characteristic total energy output of a nova is $\sim 10^{45}$ erg.

- This is about 10^{12} times more energy than the Sun produces each second.
- The duration of the thermonuclear runaway that produces most of this energy is 100-1000 seconds.
- Despite this large energy output, a nova outburst typically ejects only about 10^{-4} of the mass of the white dwarf, thus leaving the white dwarf largely intact.

This is confirmed by observation of *recurrent novae,* where after a nova the white dwarf begins accumulating accreted material again that eventually will trigger a new explosion.

RS Ophiuchi is an example of a recurrent nova.

- It consists of a *white dwarf and red giant binary*, some 5000 lightyears away from us in the constellation Ophiuchus.
- It has been observed in nova outburst *six times since 1898.*
- In its quiet phase RS Ophiuchi has an apparent visual magnitude of about $m_V \sim 12.5$.
- In nova outburst this can rise to $m_V \sim 5$.

Table 19.1: Nova Cygni 1992 abundances.

Chemical element	Abundance relative to H
He	4.5
C	70.6
N	50.0
O	80.0
Ne	250.0
Na	37.4
Mg	129.4
Al	127.5
Si	146.6
S	1.0
Ar	5.0
Ca	46.8
Fe	8.0
Ni	36.0

19.1.4 Nucleosynthesis in Novae

The hot CNO cycle leads to synthesis of new elements.

- The species of elements produced in nova explosions are relatively few in number compared with other events like supernova explosions.
- However, certain isotopes likely owe their existence primarily to nova events.

The inferred abundances of elements in the expanding shell around Nova Cygni 1992 are given in Table 19.1.

19.2 The X-Ray Burst Mechanism

In an *X-ray burst* the mechanism is thought to be similar to a nova, except that the star onto which the matter accretes is a *neutron star* rather than a white dwarf.

X-ray bursters may also produce steady X-ray emission upon which bursts are superposed.

- The steady emission is probably due to heating of matter in the accretion disk.
- Flickering is sometimes observed for the more steady emission. It is probably caused by accretion-disk instabilities.

The X-ray burst is triggered by a *thermonuclear runaway under degenerate conditions*, as for a nova.

- However, the gravitational field of a neutron star is much stronger than that of a white dwarf.
- Thus, matter falling onto the neutron star is accelerated to high velocities and
- the accretion-induced thermonuclear runaway occurs at *much higher temperatures and densities* than in a nova outburst.
- This in turn tends to produce *X-rays rather than visible light* in the thermonuclear runaway.

Production of X-rays

X-rays are emitted when fast-moving electrons pass close to slow-moving ions and are accelerated.

- Only if the *temperatures are millions of degrees* are the electrons moving at high enough velocities to produce X-rays.
- The higher the temperature, the faster the electrons move.
- This both increases the energy of the X-rays and their intensity, since collisions become more violent and more frequent at high temperature.
- An X-ray burst on the surface of a neutron star may last for a few seconds, during which time the temperatures can reach $\sim 10^9$ degrees.
- This causes X-rays to be produced in abundance.

Most nova events have maximum temperatures in the vicinity of several times 10^8 degrees, and this tends to produce light at visible and other longer wavelengths.

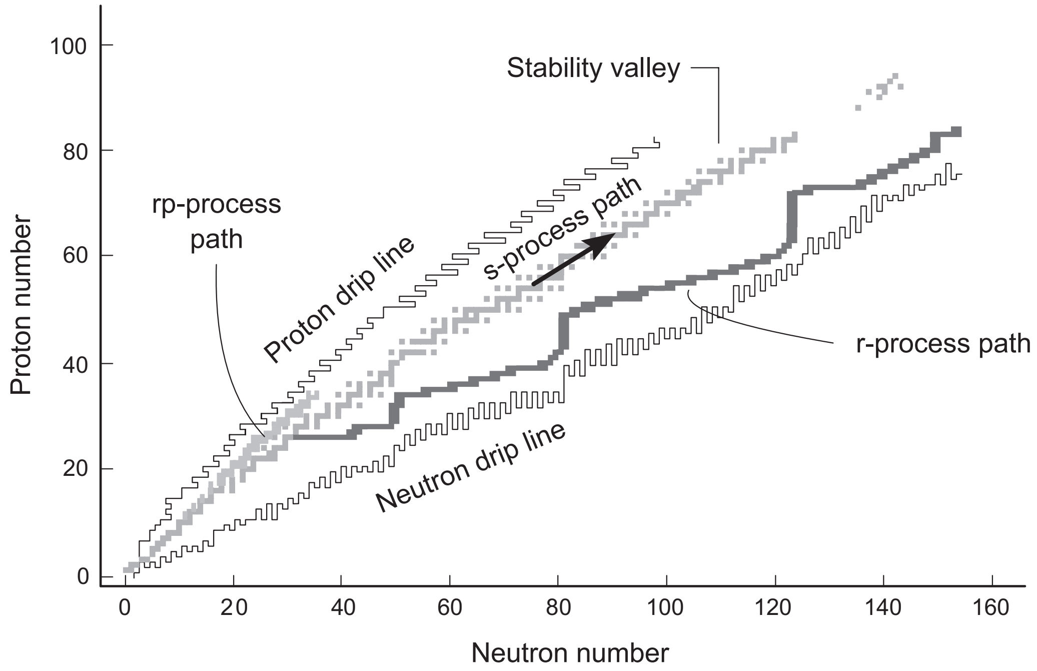
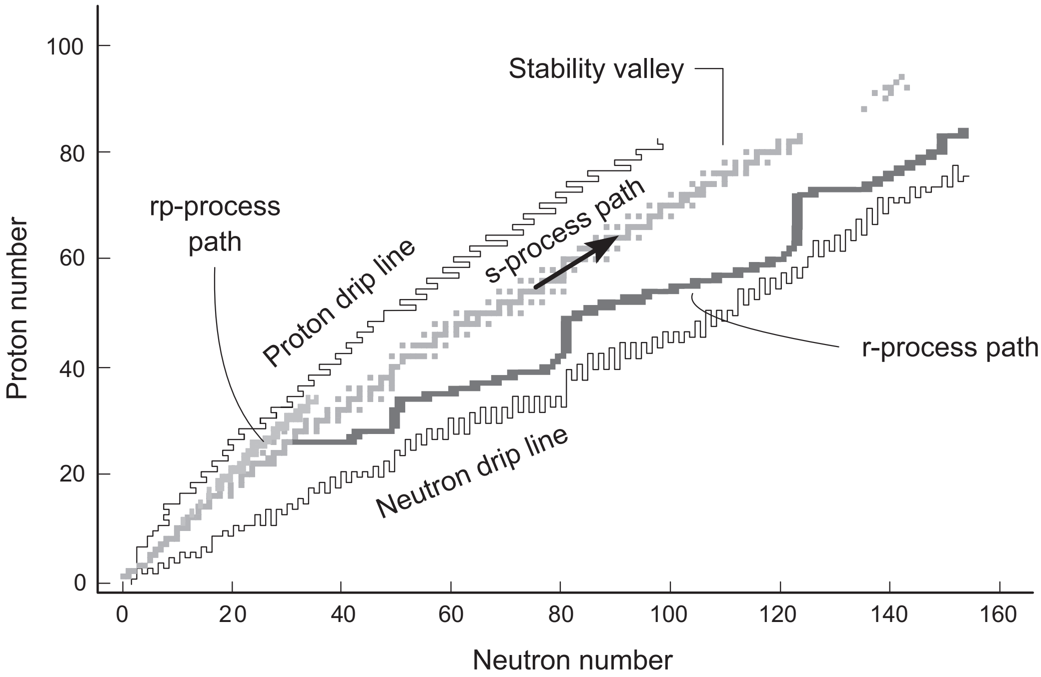


Figure 19.7: The path for the rp-process. Also shown are the s-process and r-process paths.

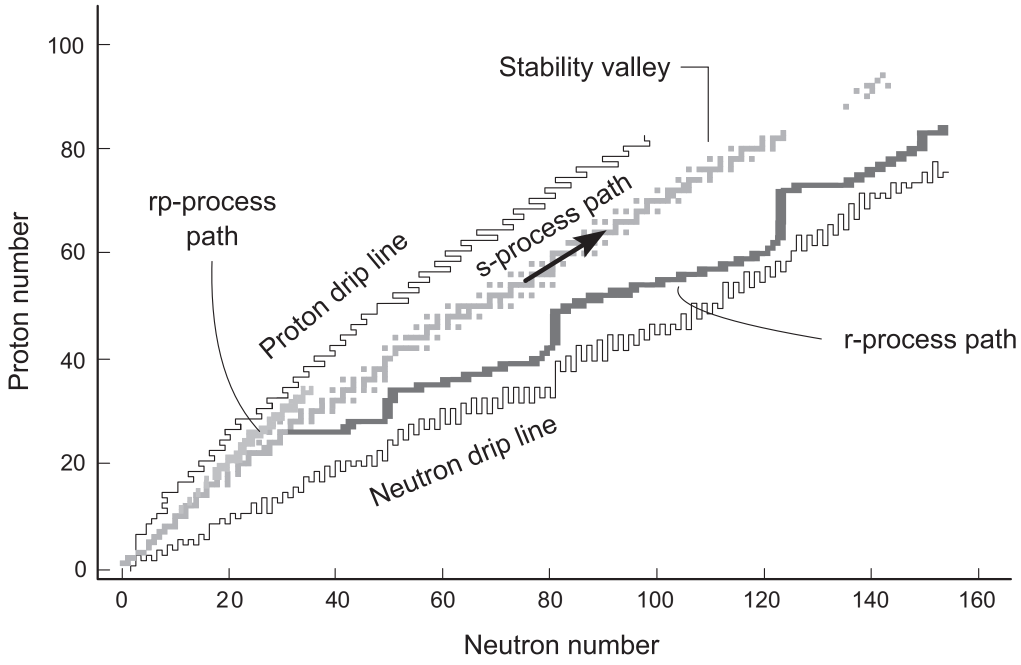
19.2.1 Rapid Proton Capture

Temperatures in an X-ray burst can become very high compared with a nova ($T > 10^9$ K is possible).

- Then the hot CNO cycle that powers novae can break out into a much more extensive network of reactions involving competition between proton capture, α-particle capture, and β-decay.
- This is called the *rapid proton capture process* or *rp-process.*
- The rp-process path is illustrated in Fig. 19.7.



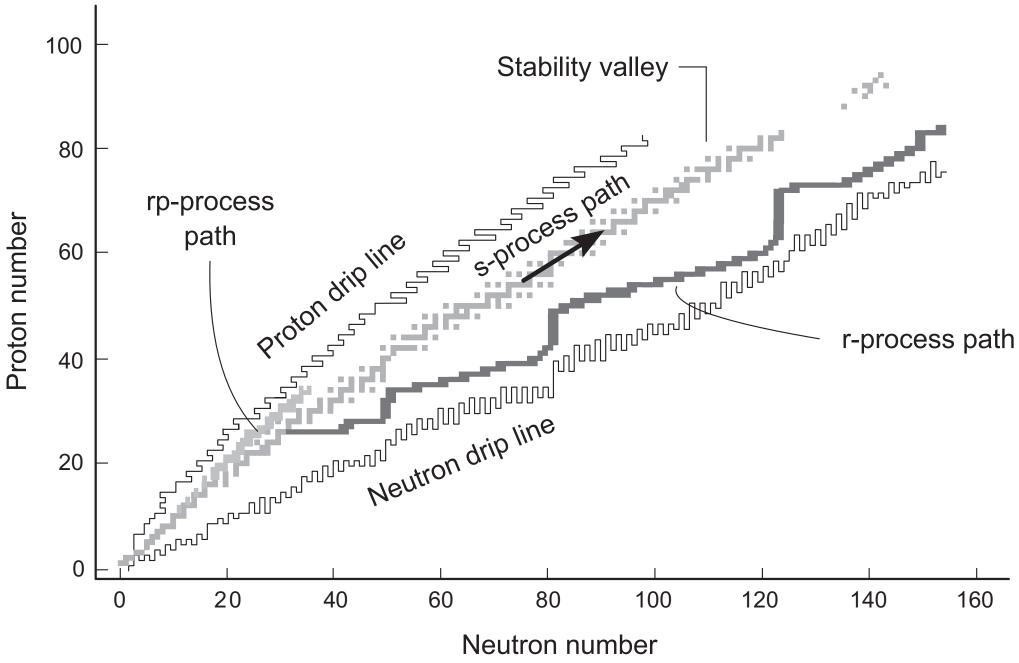
- The energy released in the reactions of the hot-CNO cycle and the rp-process are thought to provide the primary power source for X-ray bursts.
- The typical duration of the thermonuclear runaway powering the burst is a few seconds, during which time up to 10^{39} erg may be released, largely as X-rays.
- X-ray bursts are typically highly recurrent, with some repeating on timescales as short as hours.



19.2.2 Nucleosynthesis and the rp-Process

It is not known precisely how high in proton and neutron number nucleosynthesis can go during a strong X-ray burst.

- This is because of uncertainties in nuclear reaction rates and conditions characterizing a burst)
- The rp-process could be responsible for producing many of the isotopes occurring naturally that are found on the *proton-rich side* of Fig. 19.7 (isotopes lying to the left of the β-stability valley).
- A major uncertainty in this statement is whether the rp-process can lead to ejection of the synthesized elements into the interstellar medium.



- Because the gravitation field of a neutron star is so large, even if proton-rich nuclei are produced by the rp-process it is not clear that they can escape the gravity of the neutron star.
- It has been proposed that some material might escape in special circumstances, but this is uncertain.

Chapter 20

Supernovae

Supernovae represent the *catastrophic death of certain stars*. They are among the most violent events in the Universe,

- They typically *release about* 10^{53} *erg of energy*,
- much of it *in the first second* of the explosion.
- Comparison: *total luminosity of the Sun* is only about $10^{33}\,\mathrm{erg\,s^{-1}}$,
- and even a *nova outburst* releases only of order 10^{47} erg over a characteristic period of a few hundred seconds.

There is more than one type of supernova, with two general methodologies for classification:

1. According to the *spectral and lightcurve propertiesz*,
2. According to the *fundamental explosion mechanism* responsible for the energy release.

In addition to their intrinsic interest, supernovae of various types are of fundamental importance for a variety of astrophysical phenomena, including

- *element production* and galactic chemical evolution,
- potential *relationship to some types of gamma-ray bursts*,
- *energizing and compressing the interstellar medium* (implying a connection with star formation),
- *gravitational wave emissio*n, and
- *distance-measuring applications in cosmology* associated with *standardizable candle properties*.

We begin our discussion by considering the taxonomy of these events.

20.1 Classification of Supernovae

The traditional classification of supernovae is based on observational evidence, primarily their

- spectra
- lightcurves.

Some *representative supernova spectra* are displayed in Fig. 20.1 on the following page.

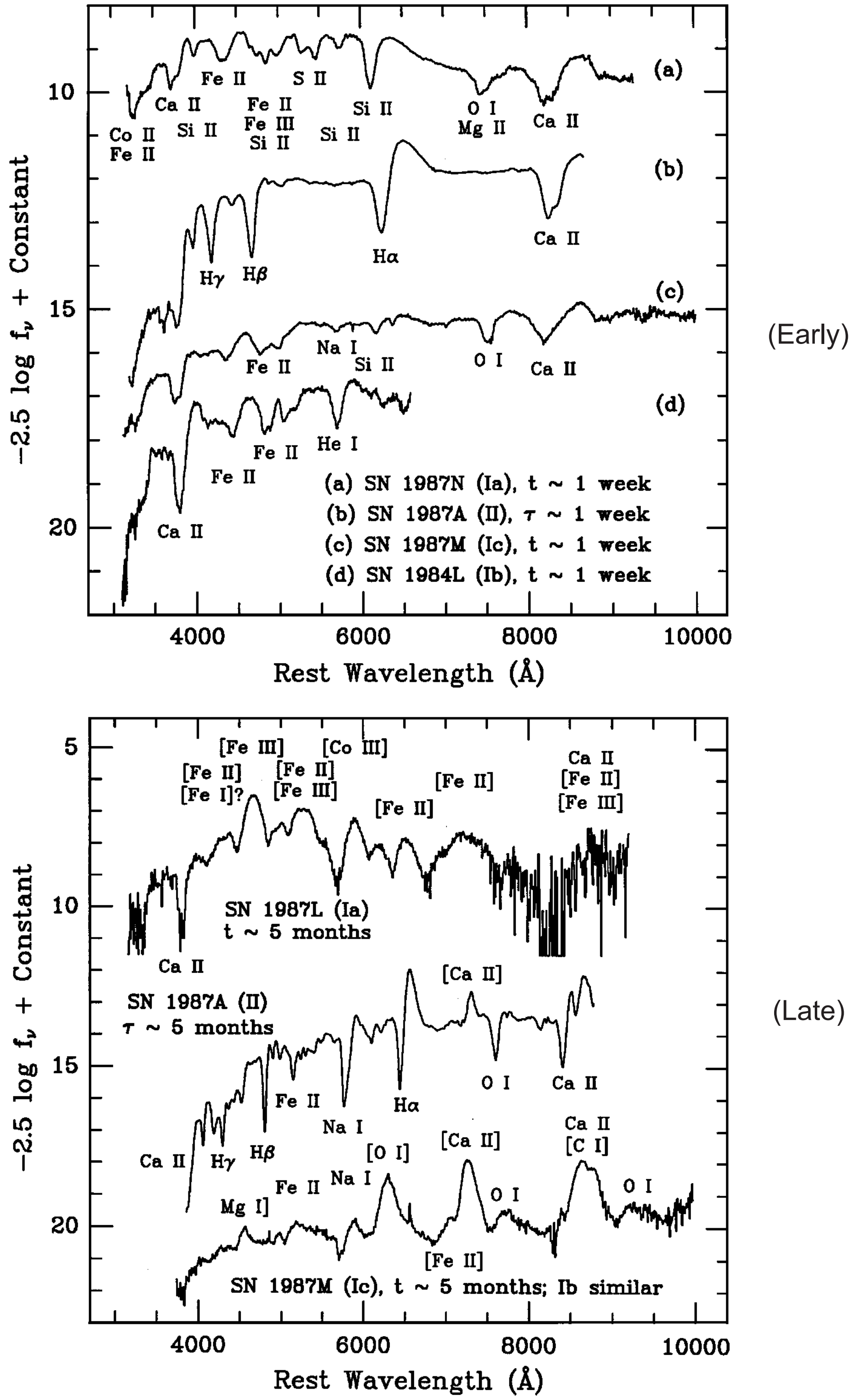


Figure 20.1: *Early-time and late-time spectra* for several classes of supernovae.

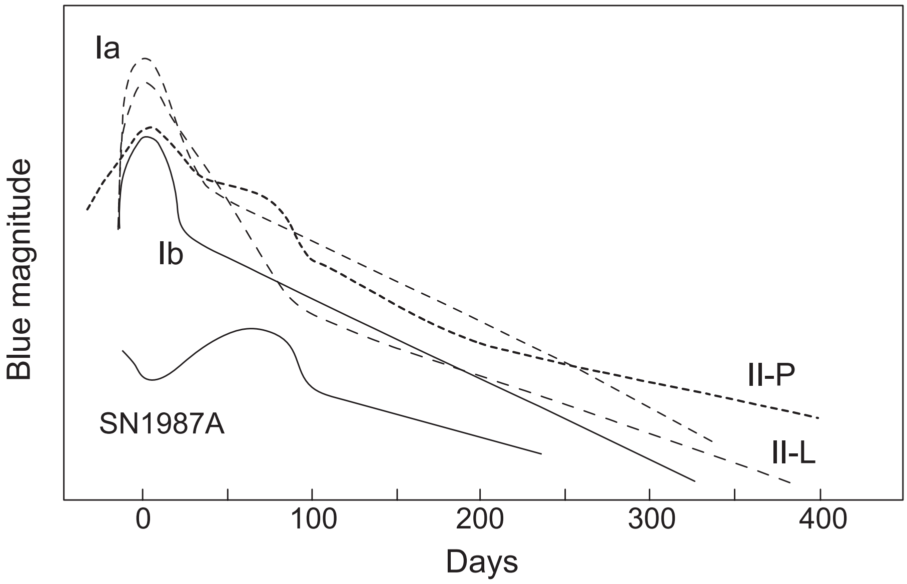


Figure 20.2: *Schematic lightcurves* for different classes of supernovae.

Some typical lightcurves are illustrated in Fig. 20.18.

- In most cases now we have at least a schematic model that can be associated with each class
- that can account for the observational characteristics of that class.
- Those models suggest that all supernova events derive their enormous energy from
  - *gravitational collapse* of a massive stellar core or
  - a *thermonuclear runaway* in dense, electron-degenerate matter.

The observational characteristics of supernovae derive both from

- the *internal mechanism causing the energy release* (for example, collapse of a stellar core) and
- the *interaction of the energy release with the surrounding medium* (outer layers or extended atmosphere of the star).

Therefore

- some observational characteristics are direct *diagnostics of the explosion mechanism itself*,
- while others are only indirectly related to the explosion mechanism and instead are *diagnostics for the state of the star and its surrounding medium* at the time of the outburst.

The *standard classes of supernovae* and some of their characteristics are illustrated in Fig. 20.3.

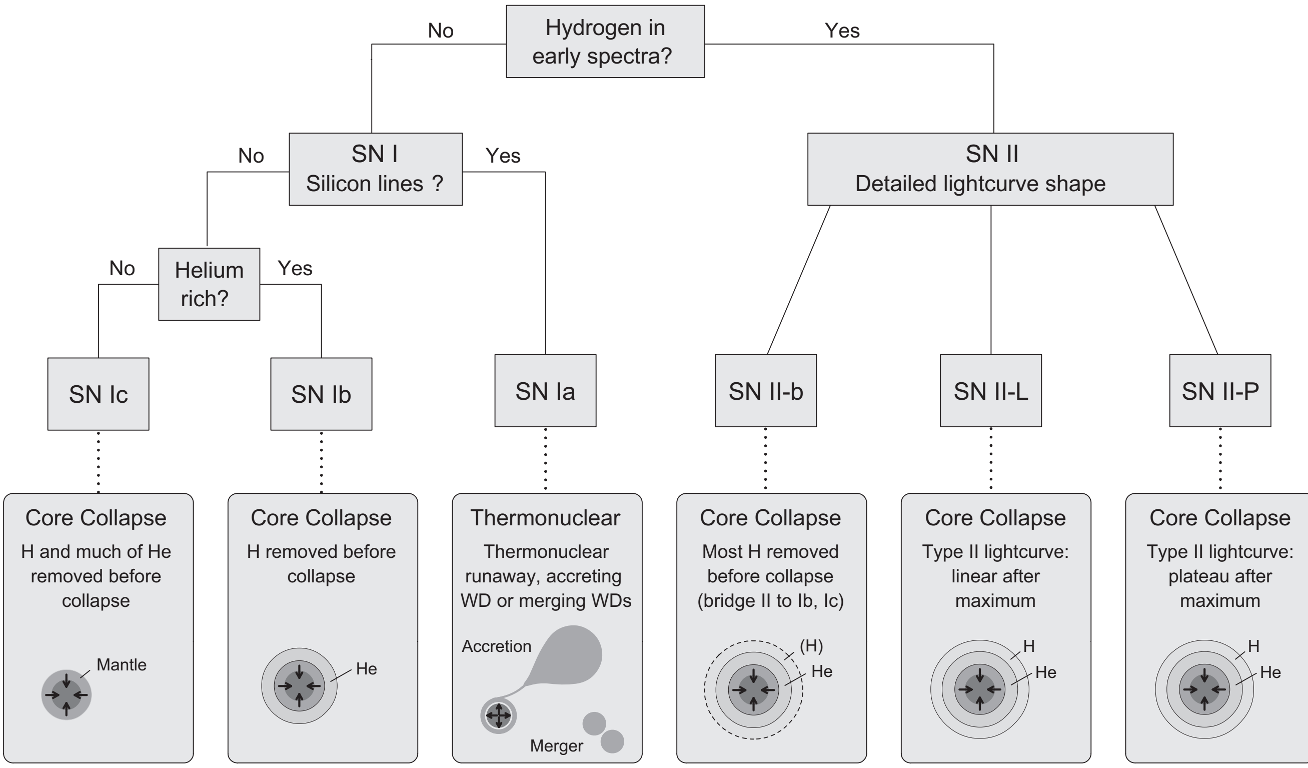

Hydrogen in early spectra?
No
Yes
SN I
Silicon lines ?
No
Yes
SN II
Detailed lightcurve shape
Helium rich?
No
Yes
SN Ic
SN Ib
SN Ia
SN II-b
SN II-L
SN II-P
Core Collapse
H and much of He removed before collapse
Mantle
Core Collapse
H removed before collapse
He
Thermonuclear
Thermonuclear runaway, accreting WD or merging WDs
Accretion
Merger
Core Collapse
Most H removed before collapse (bridge II to Ib, Ic)
(H)
He
Core Collapse
Type II lightcurve: linear after maximum
H
He
Core Collapse
Type II lightcurve: plateau after maximum
H
He

Figure 20.3: *Classification of supernova events*. Note that Type II-n is not shown here but discussed in the text.

The primary initial distinction concerns *whether hydrogen lines are present* in the spectrum, which divides supernovae into

- *Type I* (no hydrogen lines) and
- *Type II* (significant hydrogen lines).

The standard subclassifications then correspond to the following characteristics:

20.1.1 Type Ia

A *Type Ia supernova* is thought to be associated with a *thermonuclear runaway under degenerate conditions in white dwarf matter*.

- This class of supernovae is sometimes termed a thermonuclear supernova.
- This distinguishes it from all other classes that derive their power from gravitational collapse and not from thermonuclear reactions.
- No hydrogen is observed but calcium, oxygen, and silicon appear in the spectrum near peak brightness.
- Type Ia supernovae are found in all types of galaxies and their standardizable candle properties make them a valuable distance-measuring tool (see following Box).

Standard and Standardizable Candles

We may distinguish

- A *standard candle*, which is a light source that always has the same intrinsic brightness under some specified conditions.
- A *standardizable candle*, which is a light source that may vary in brightness but that can be standardized (normalized to a common brightness) by some reliable method.

Standard candles, or standardizable candles, then permit distance measurement by *comparing observed brightness with the standard brightness*.

- Different Type Ia supernovae have similar but not identical lightcurves.
- Hence they are *not standard candles*.
- However, there are empirical methods that allow the lightcurves of different Type Ia supernovae to be collapsed to a single curve.
- Thus, *they are standardizable candles*.

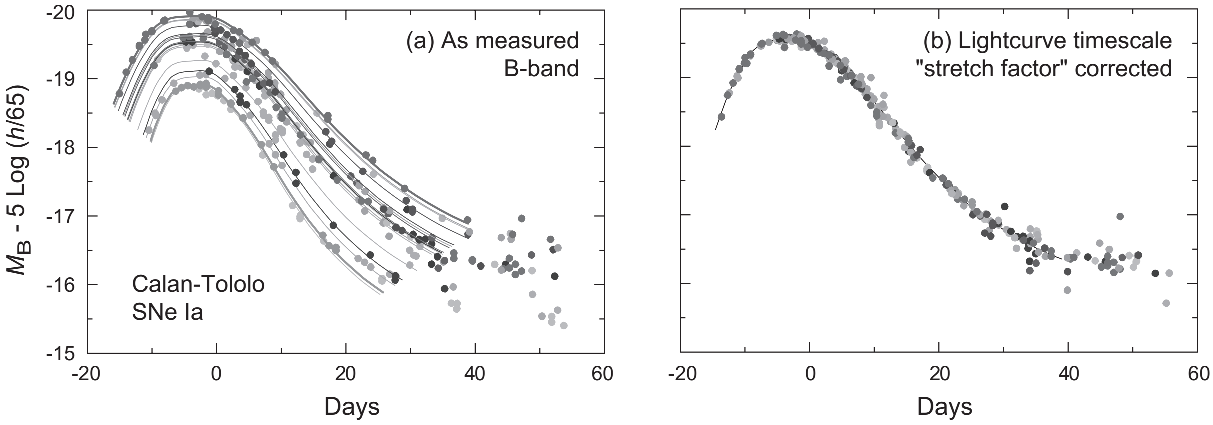


Figure 20.4: Empirical rescaling of Type Ia supernova lightcurves to make them standardizable candles. (a) B-band lightcurves for low-redshift Type Ia supernovae (Calan-Tololo survey). As measured, the intrinsic scatter is 0.3 mag in peak luminosity. (b) After 1-parameter correction the dispersion is 0.15 mag.

The establishment of Type Ia supernovae as *standardizable candles* is illustrated in Fig. 20.4.

Type Ia standardizable candles are particularly valuable because their *extreme brightness* makes them *visible at very large distances*.

- The standardizable candle and brightness properties of Type Ia supernovae have made them *central tools of modern cosmology*.
- For example, they are the *most direct indicator of accelerated expansion of the Universe*.
- This implies that the Universe is permeated by a mysterious *dark energy* that can effectively turn *gravity into anti-gravity*.

20.1.2 Type Ib and Type Ic

Type Ib and Ic supernovae are thought to represent

- core collapse of a massive star that has
- lost much of its outer envelope because of strong stellar winds or interactions with a binary companion.

These progenitors are called *Wolf–Rayet stars*.

The distinction between Types Ib and Ic is thought to lie in whether

- only the hydrogen envelope has been lost before core collapse *(Type Ib)*, or
- whether most of the helium layer has also been expelled *(Type Ic)*.

20.1.3 Type II

Type II supernovae are characterized by prominent hydrogen lines.

- They are thought to be associated with the core collapse of a massive star.
- They are found only in regions of active star formation.

For example, Type II supernovae are unlikely to be found in elliptical galaxies.

Type II supernovae may be further subdivided according to detailed *spectral and light-curve properties:*

1. ***Type II-P***: In the designation II-P, the P refers to a plateau in the light curve.

2. ***Type II-L***: In the designation Type II-L, the L refers to a linear decrease of the light curve in the region where a Type II-P lightcurve has a plateau.

3. ***Type II-b***: In a Type II-b event the spectrum contains prominent hydrogen lines initially, but it transitions into one similar to that of a Type Ia,b supernova.

   > The suspected II-b mechanism is core collapse in a red giant that has lost most but not all of its hydrogen envelope through stellar winds or interaction with a binary companion.

4. ***Type II-n***: In this class of core-collapse supernova, narrow emission lines and a strong hydrogen spectrum are present.

   > Type II-n supernovae are thought to originate in the core collapse of a massive star embedded in dense shells of material ejected by the star shortly before the explosion.

We conclude that

- all of the Type-II subcategories, and
- the Type Ib and Type Ic subcategories,

correspond to a similar core-collapse mechanism. The observational differences derive primarily from

- differences in the outer envelope and
- their influence on the spectrum and lightcurve,

not in the primary energy-release mechanism.

20.2 Type Ia (Thermonuclear) Supernovae

A Type Ia supernova is thought to correspond to

- a *runaway thermonuclear explosion* that occurs in electron-degenerate, carbon–oxygen white dwarf matter
- that is triggered by
  - *accretion on a white dwarf* from acompanion that isn't a white dwarf, or
  - *merger* of two white dwarfs

  in a binary system.

Thus it differs fundamentally from all of the other classes of supernovae.

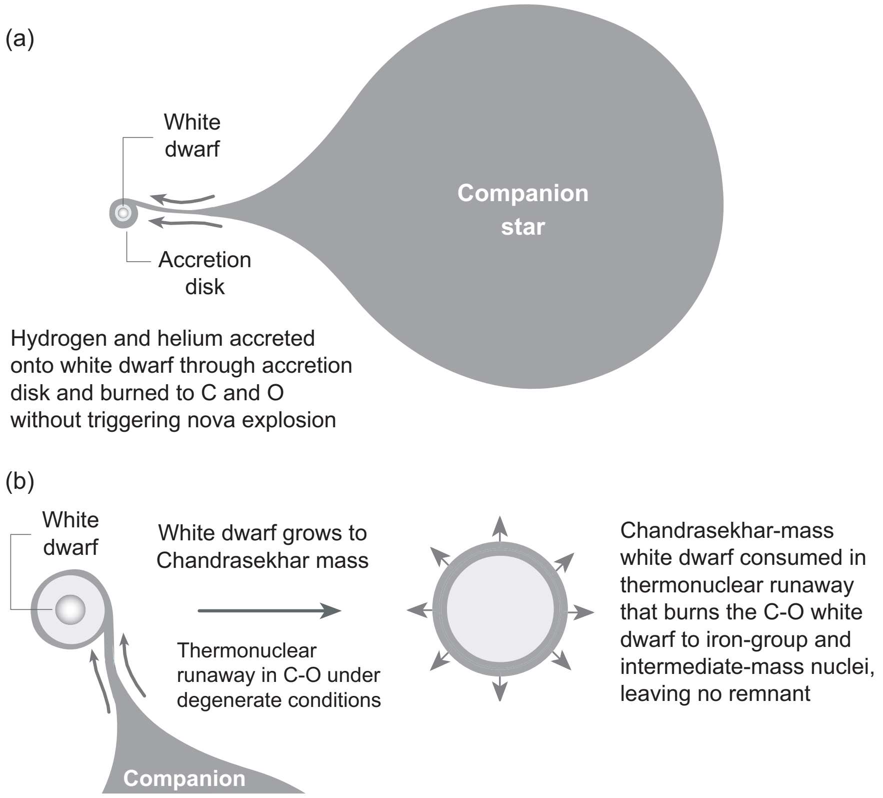
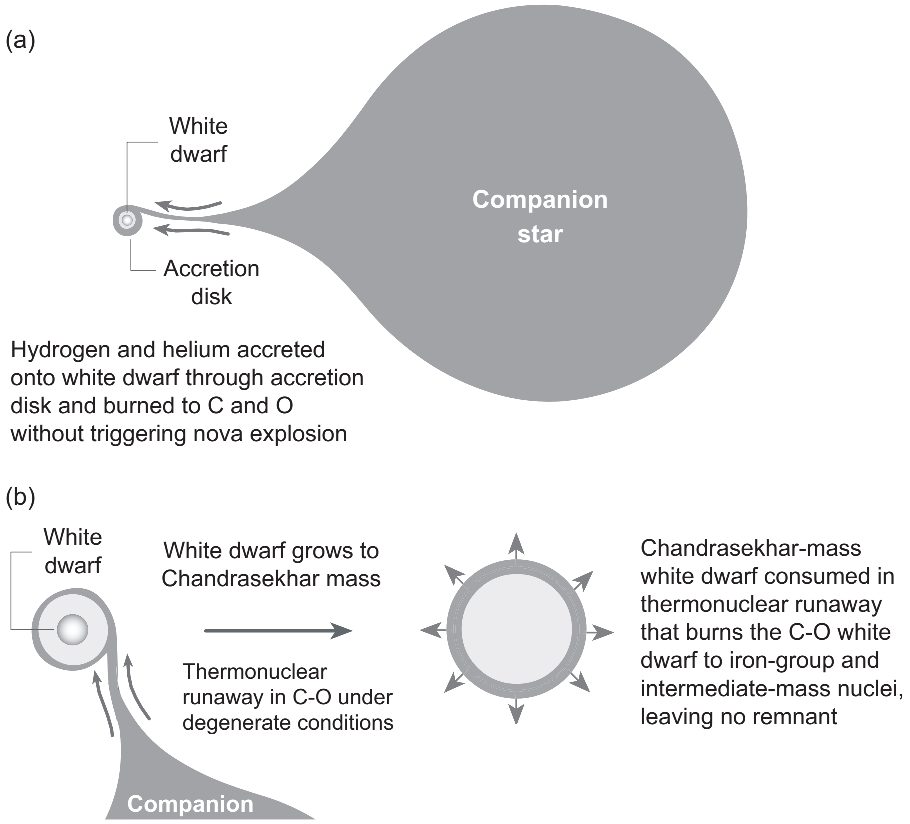

(a)
White dwarf
Companion star
Accretion disk
Hydrogen and helium accreted onto white dwarf through accretion disk and burned to C and O without triggering nova explosion
(b)
White dwarf
White dwarf grows to Chandrasekhar mass
Thermonuclear runaway in C-O under degenerate conditions
Chandrasekhar-mass white dwarf consumed in thermonuclear runaway that burns the C-O white dwarf to iron-group and intermediate-mass nuclei, leaving no remnant
Companion

Figure 20.5: The single-degenerate mechanism for a Type Ia supernova.

20.2.1 The Single Degenerate Mechanism

One Type Ia scenario is illustrated in Fig. 20.5. It is related to the nova scenario. The difference is that

- In a nova a thermonuclear runaway is initiated in a thin surface layer after some accretion and the white dwarf remains largely intact after the explosion.
- In the Type Ia situation the matter accumulates on the surface of the white dwarf over a long period without triggering a runaway in the accumulated surface layers.


(a)
White dwarf
Accretion disk
Companion star
Hydrogen and helium accreted onto white dwarf through accretion disk and burned to C and O without triggering nova explosion
(b)
White dwarf
White dwarf grows to Chandrasekhar mass
Thermonuclear runaway in C-O under degenerate conditions
Chandrasekhar-mass white dwarf consumed in thermonuclear runaway that burns the C-O white dwarf to iron-group and intermediate-mass nuclei, leaving no remnant
Companion

- As unburned matter accumulates, the mass of the white dwarf grows and can approach the Chandrasekhar limit.
- Near the Chandrasekhar limit the very high density can trigger a thermonuclear runaway in the interior of the star that initially ignites carbon and then oxygen.
- Quickly (in a matter of a second or less) the runaway burns a large percentage of the mass of the white dwarf to iron-group nuclei, with an enormous release of energy
- Most of the iron in the Universe probably originates in Type Ia and core-collapse, supernovae.

Thus, unlike

- a nova, or
- a core-collapse supernova,

a Type Ia supernova does not leave behind a compact remnant

Single-Degenerate and Double-Degenerate Scenarios

The Type Ia mechanism described above is sometimes termed the *“singly-degenerate model”*, since it involves only one degenerate object (the white dwarf).

- Alternative mechanism: triggering of a thermonuclear burn by merger of two white dwarfs.
- This is called the *“double-degenerate” model*, because it involves two degenerate objects.
- In either the single-degenerate or double-degenerate models the cause of the explosion is the same: *A thermonuclear runaway in dense electron-degenerate matter*.
- The primary difference is in how the explosion is initiated.
- At present neither model can yet describe all aspects of a Type Ia explosion without assumptions.
- It is possible that more than one progenitor scenario is needed to explain all Type Ia supernovae.

One side benefit of the singly-degenerate model is that it can provide a plausible explanation for the standardizable candle property.

- The Chandrasekhar mass is almost the same for all white dwarfs.
- Thus, if the white dwarf that explodes is always near the Chandrasekhar mass it makes sense that the total energy produced by different Type Ia events is similar.
- In contrast, for the doubly-degenerate model there is no obvious reason for the sum of the masses of the two white dwarfs that merge to be similar in different events.

However, the preceding may be an oversimplified analysis.

- The later-time Type Ia lightcurve is largely determined by how much ^{56}Ni is produced in the explosion.
- Thus the standardizable candle property could result from any mechanism that causes a similar amount of ^{56}Ni to be made in all Type Ia explosions.

20.2.2 Thermonuclear Burning under Extreme Conditions

Because of the gigantic energy release in a small region over a very short period of time, the conditions in a Type Ia explosion are extreme. Simulations indicate that

- Temperatures in the hottest parts can approach 10^{10} K, with densities as large as $10^9\,\mathrm{g\,cm^{-3}}$.
- In the thermonuclear burn front, temperature changes as large of 10^{17} K/s are seen in simulations.

The physics of the Type Ia explosion presents a number of issues that are difficult to deal with in the large numerical simulations that are required to model such events, as discussed in the following box.

Thermonuclear Burns: Different Scales

In the Type Ia explosion a thermonuclear burn corresponding to conversion of carbon and oxygen fuel into heavier elements by nuclear reactions releases large amounts of energy.

- This burn is extremely violent and involves energy and temperature scales far beyond our everyday experience.
- But it shares many qualitative properties with ordinary chemical burning.
- There is a *burn front* that proceeds through the white dwarf, with "cooler" (a highly relative term!) unburned fuel in front and hot burned products (ash) behind.
- This burn front can be remarkably narrow—as small as millimeters.

Thus there are two extremely different distance scales characterizing the explosion:

- the *size of the white dwarf*, which is of order 10^4 km, and
- the *width of the burn front* that consumes it, which can be *billions of times smaller*.

> This presents severe difficulties in accurately modeling Type Ia explosions, since standard numerical methods to solve the equations cannot handle such disparate scales without drastic approximation.

DeFlagration and Detonation Waves

In thermonuclear and ordinary chemical burning there is an important distinction associated with the speed of the burn front.

- If the burn front advances through the fuel at a speed less than the speed of sound in the medium *(subsonic)*, it is termed a *deflagration wave.*
- If the burn front advances at greater than the speed of sound in the medium *(supersonic)* it is called a *detonation wave.*

Deflagration and detonation waves have different characteristics:

- In a deflagration, fuel in front of the advancing burn is heated to the ignition temperature by *conduction of heat across the burn front*.

> Recall that matter described by a degenerate equation of state is a *very good thermal conductor*, much like a metal.

- In a detonation a shock wave forms and the fuel in advance of the burn front is brought to ignition temperature by *shock heating*.
- Generally *detonation is more violent than deflagration*.

DeFlagration, Detonation and Isotopic Abundances

Deflagrations and detonations produce different isotopic abundance signatures in the ash that is left behind.

- The elemental abundances detected in the expanding debris of Type Ia explosions could be accounted for most naturally if we assume that
  - part of the burn is a deflagration and
  - part of it is a detonation.
- This is a difficulty for the theory because general considerations suggest that
  - the explosion starts off as a deflagration and
  - it is not easy to get the burn in computer simulations to transition to a detonation without making significant untested assumptions.

Thus, we believe that the proposed Type Ia mechanism is plausible in outline, but there are bothersome details that leave some doubt about whether we understand fully the mechanism of these gigantic explosions.

20.2.3 Element and Energy Production

The energy released in a Type Ia supernova explosion derives primarily from the thermonuclear burning of carbon and oxygen to heavier nuclei.

- If the explosion lasts long enough to achieve nuclear statistical equilibrium (NSE), the primary final products of this burning will be iron-group nuclei.
- An example of network evolution under conditions typical of the Type Ia explosion in the deep interior of the white dwarf is illustrated in Fig. 20.6 (next page).

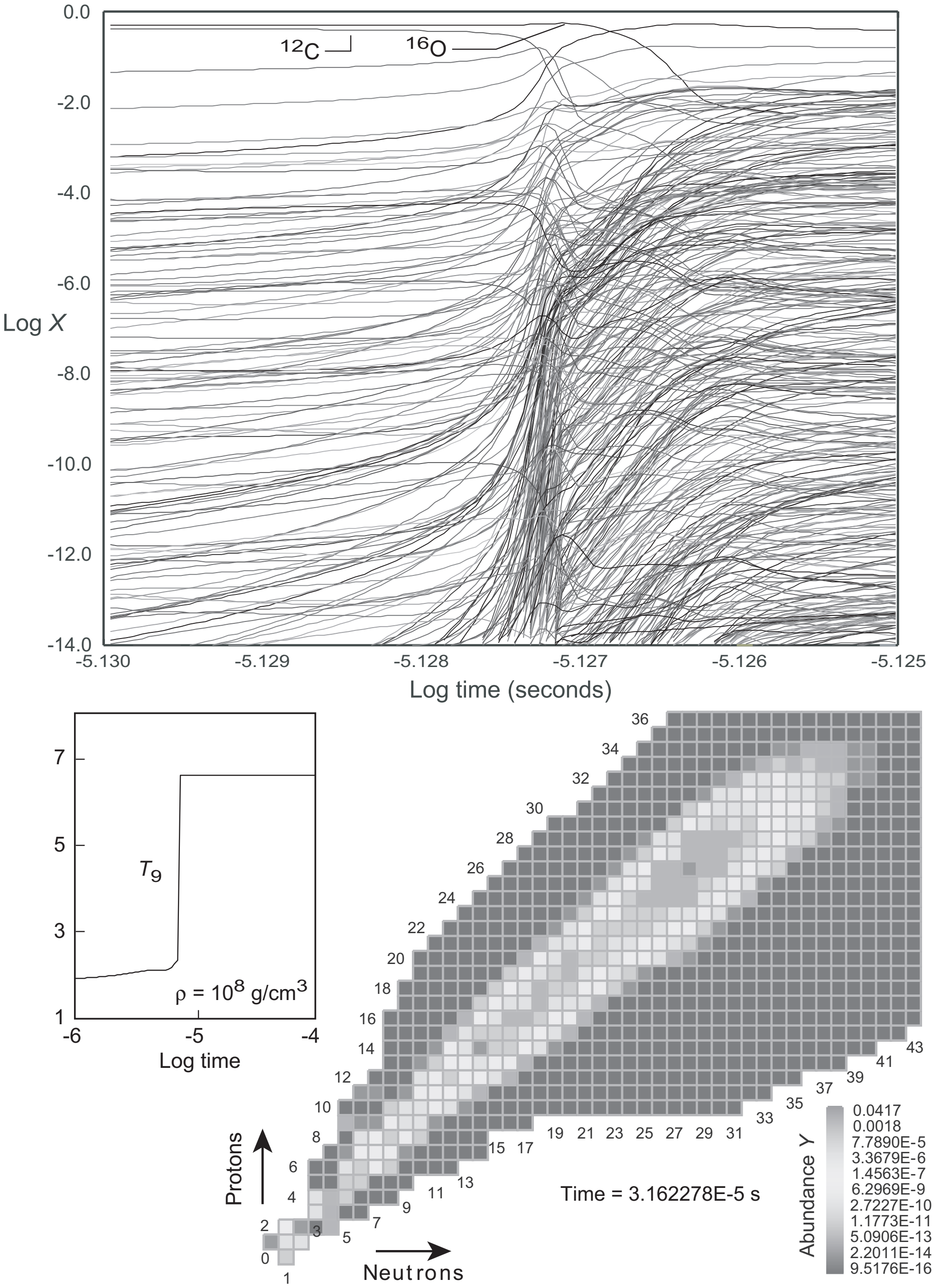


Figure 20.6: Element production in a Type Ia explosion. Upper: mass fractions X for 468 isotopes and integrated energy production $\sum E$ in MeV per nucleon. Lower: distribution of abundances Y for all isotopes at end of calculation in the upper figure. Inset on left shows the variation of temperature with time (density remains almost constant over this time period).

- In this calculation
  - the initial temperature was $T_9 = 2$,
  - the initial density was $\rho = 1 \times 10^{-8}\,\mathrm{g\,cm^{-3}}$, and
  - the initial composition was equal mass fractions of ^{12}C and ^{16}O.

- The explosion is *initiated by carbon burning*, which quickly raises the temperature (see the inset to the figure) and initiates burning of oxygen and all the reaction products that are produced by carbon and oxygen burning.

- The rapid temperature rise is associated with the coupling of the large energy release from the thermonuclear burning described by the reaction network to the fluid of the white dwarf, which is described by hydrodynamics.

- This energy release (through the equation of state) causes a rapid rise in temperature of the fluid representing the white dwarf.

- This in turn increases rapidly the rate of nuclear reactions in the network.

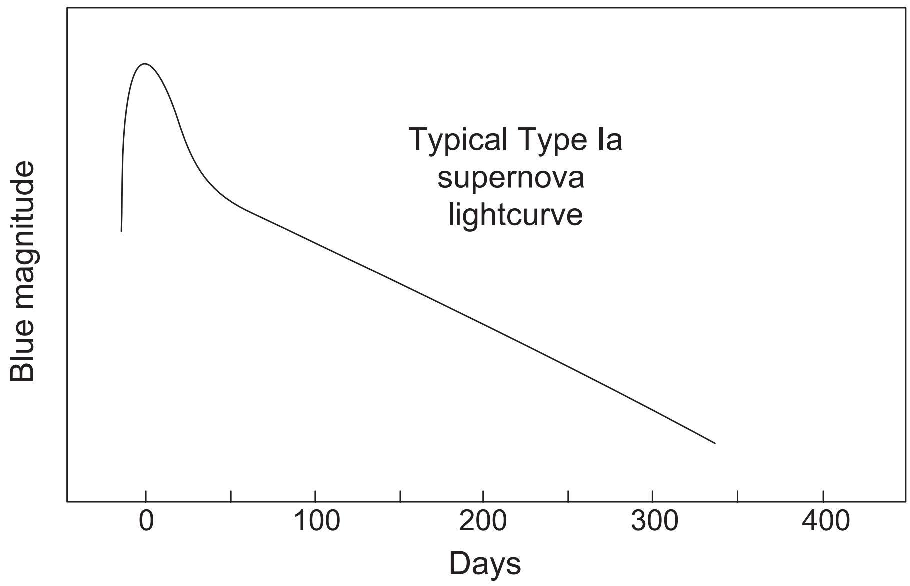
- The net result is the almost vertical rise in temperature from $T_9 \sim 2$ to $T_9 \sim 6.6$ in a period of less than 10^{-5} s.

- During this time the number of isotopic species in the network has increased from *two to about five hundred*.

- *Significant population of iron group nuclei* is already evident.

Under these conditions, as the thermonuclear flame burns through the white dwarf

- the carbon and oxygen fuel in each region is burned in a tiny fraction of a second, and
- the entire white dwarf is consumed by the thermonuclear flame on a timescale of less than a second.



20.2.4 Late-Time Observables

A Type Ia supernova is expected to leave no remnant behind. The primary late-time observables are

- the *supernova lightcurve* and
- the *motion and spectrum* of the expanding *supernova remnant*.

A typical shape of a Type Ia lightcurve is shown above.

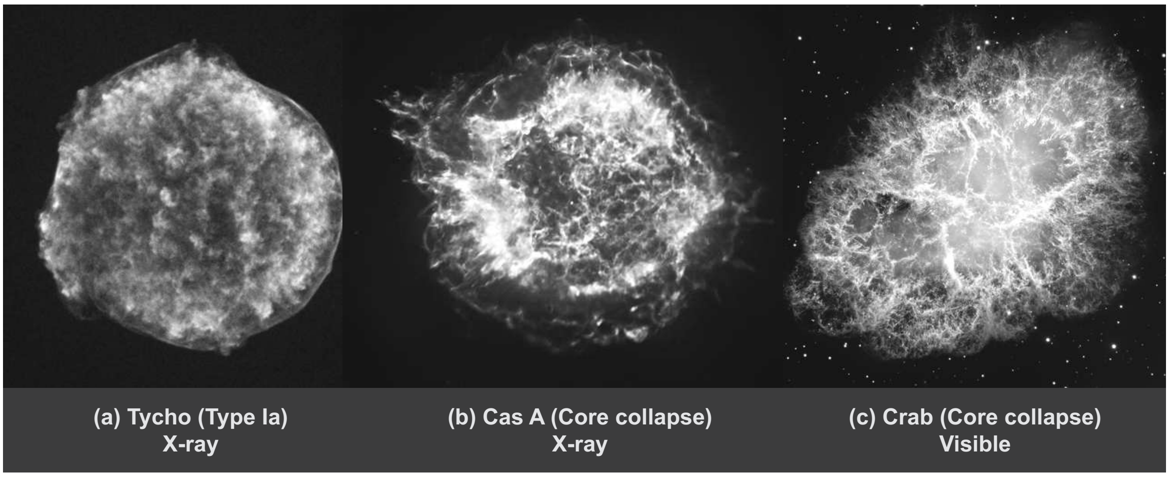


Figure 20.7: Supernova remnants. (a) Tycho's supernova of 1572 in X-rays; it was a Type Ia. (b) Cas A supernova remnant in X-rays. It was a core collapse supernova that occurred about 300 years ago. (c) Crab Nebula, which is the remains of the core collapse supernova of 1054, in visible light.

Some supernova remnants are shown in Fig. 20.7.

- Figure (a) is an example of a *Type Ia remnant*.
- Spectroscopy of a Type Ia remnant can determine
  - the elements in the debris and
  - their radial velocity.
- The radial velocity is in turn correlated with how deep in the explosion the element was produced (higher velocity is expected to come from deeper).
- Measurements indicate that *many intermediate-mass species like Si are produced*; not just iron-group nuclei.

This suggests that the explosion is not a pure detonation (which would produce mostly iron-group nuclei).

Another proposed supernova mechanism involves some aspects of both core collapse and thermonuclear runaway: massive stars of low metallicity can undergo a *pair-instability supernova.*

20.3 Pair-Instability Supernovae

More *massive stars* ($M \sim 130 - 250\,M_\odot$) of *low metallicity* are predicted theoretically to undergo a *pair-instability supernova.*

- In very massive stars the *radiation pressure* is primarily responsible for balancing the enormous gravity, with the gas pressure playing a smaller role.
- At high temperatures and densities *energetic photons can produce electron–positron pairs* in abundance.
- This removes photons and part of the pressure support for the core.
- If pairs are produced at a high-enough rate *the core begins to collapse*,
- which leads to *increased pair production* that *accelerates the collapse*.
- This in turn *greatly accelerates thermonuclear burning* and leads to
- a *thermonuclear runaway* that blows the star apart, without leaving behind a neutron star or black hole.
- For massive progenitors, pair-instability supernovae can be brighter than Type Ia or core collapse supernovae.

> It has been proposed that some overly-luminous supernovae may have been of this type.

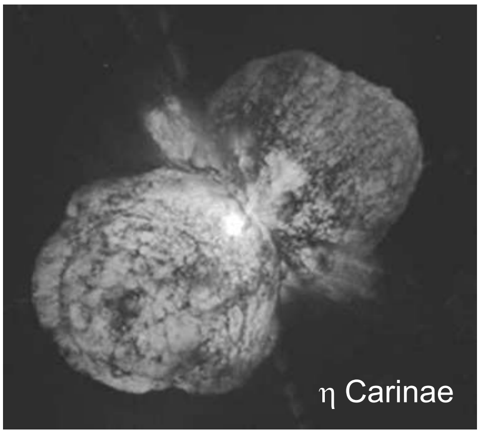
- For stars with $\mathrm{M} < 130\, M_\odot$ the *pair production rate is not high enough* to trigger the above-mentioned runaway.
- A pair-instability explosion also is unlikely *if the metallicity of the star is too high*, because
  - this *leads to high photon opacity* and
  - *prevents the runaway collapse* that initiates the explosion.
- For stars more massive than $\sim 250\, M_\odot$,
  - *photodisintegration of nuclei* removes pressure support so rapidly that
  - the star collapses to a black hole

  before encountering the pair instability.

- For stars in the mass range $\sim 100-130 M_\odot$ the pair instability does not lead to a supernova.
- However, the pair instability
  - *destabilizes the star* sufficiently that
  - it exhibits *pulsations*

  causing *large mass ejection*.

A possible explanation for the strange behavior of η Carinae:



is mass ejection caused by such an instability.

20.4 Core-Collapse Supernovae

A core-collapse supernova is one of the most spectacular events in nature, and is one possible source of the heavy elements that are produced in the rapid neutron capture or r-process.

- Considerable progress has been made in understanding the mechanisms responsible for such events.
- This understanding was tested both qualitatively and quantitatively by the observation of *Supernova 1987A* in the nearby Magellanic Cloud.
- This was the brightest supernova observed from earth since the time of Kepler and
- the first nearby supernova to occur in the era of modern scientific instrumentation.

20.4.1 The "Supernova Problem"

The observations of Supernova 1987A, and the continuing studies of its aftermath,

- Provide compelling evidence that a core-collapse supernova represents the *death of a massive star* in which
  - a *degenerate iron core* of approximately 1.2–1.3 $M_\odot$
  - *collapses catastrophically* on timescales of tens of milliseconds.
- This gravitational collapse is reversed as the inner core exceeds nuclear densities because of the properties associated with the stiff nuclear equation of state.
- A pressure wave reflects from the center of the star and propagates outward.
- The pressure wave steepens into a shock wave as it passes into increasingly less dense material of the outer core.
- This shock blows off the outer layers of the star, producing the spectacular explosion seen in observations.

However, the most realistic simulations of this event indicate that

- The shock wave loses energy rapidly as it propagates through the outer core.
- If the core has a mass of more than about $1.1\ M_\odot$,
- the shock stalls into an accretion shock within several hundred milliseconds of the bounce, several hundred kilometers from the center.
- Thus the "prompt shock" does not blow off the outer layers of the star and fails to produce a supernova.

This is the *"supernova problem":*

- there is good evidence that we understand the basics, but
- the details fail to work robustly.

In this chapter we summarize the present status of this problem.

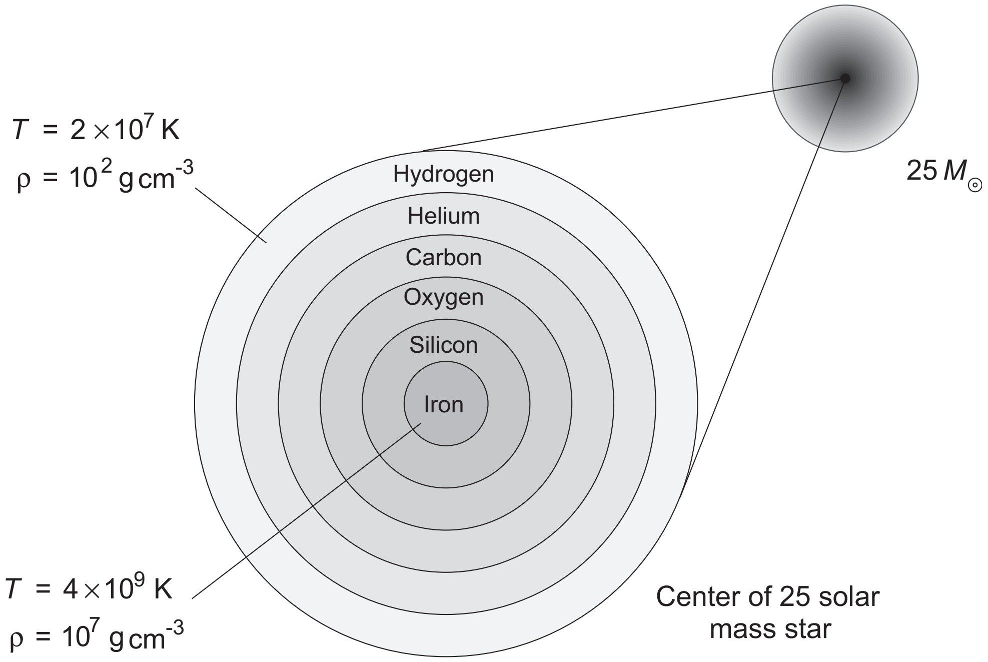
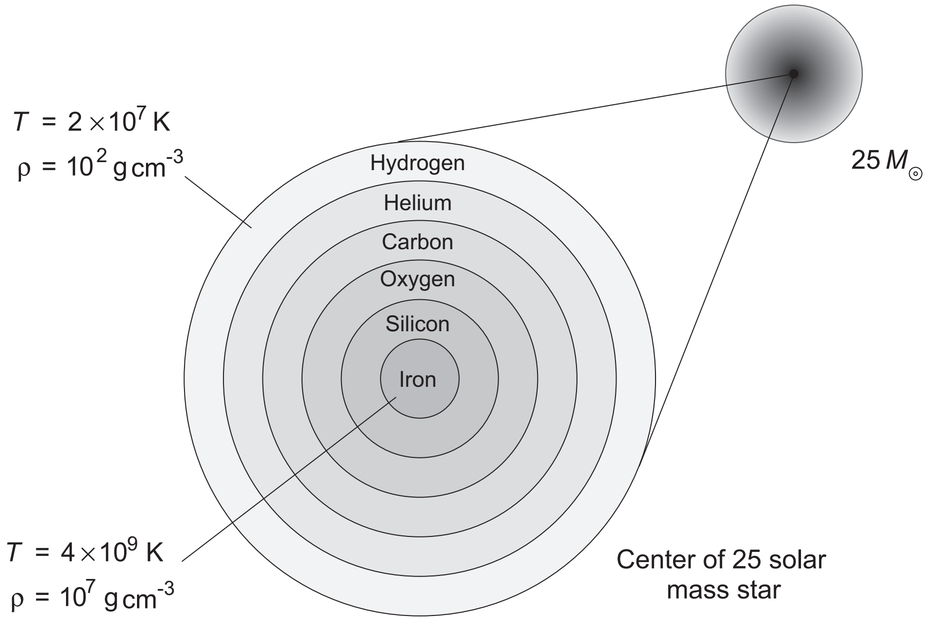


Figure 20.8: Center of a 25 solar mass star late in its life.

20.4.2 The Death of Massive Stars

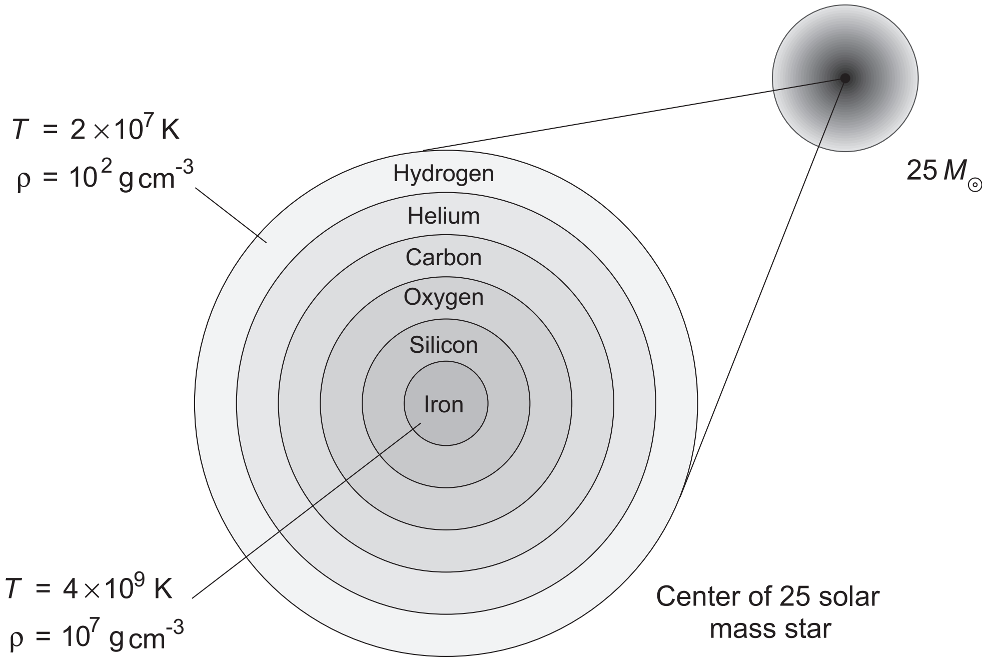
Because of sequential advanced burning stages, massive stars near the ends of their lives build up the layered structure depicted in Fig. 20.8.

- The iron core that is produced in the central region of the star cannot produce energy by fusion (the curve of binding energy peaks in the iron region).
- Thus, it must ultimately be supported by electron degeneracy pressure.
- As the silicon layer undergoes reactions the central iron core becomes more massive.



- Electron degeneracy supports the iron core against gravitational collapse only if its mass remains below the Chandrasekhar limit, which depends on the electron fraction but is approximately $1.1–1.4\,M_\odot$.
- When the core exceeds this mass it begins to collapse because electron degeneracy can no longer balance gravity.
- When the collapse begins, the core of a $25\,M_\odot$ star has a mass of about $1.2\,M_\odot$ and a diameter of several thousand kilometers.

> This is a tiny fraction of the total diameter. The star is typically a supergiant and its outer layers would encompass much of the inner Solar System if placed at the location of the Sun.



- The core density is about 6×10^9 g cm^{-3},
- the core temperature is approximately 6×10^9 K, and
- the entropy per baryon per Boltzmann constant of the core is about 1, in dimensionless units.
- This is a very small entropy, as discussed on the following page.

Entropy of the Iron Core

This entropy of the iron core in a pre-supernova star is remarkably low: the entropy of the original main-sequence star that produced this iron core is about 15 in units of entropy per baryon per Boltzmann constant.

- At first glance, it may seem contradictory for the entropy to decrease as the star burns its fuel.
- However, the star is not a closed system: as nuclear fuel is consumed, energy leaves the star in the form of photons and neutrinos.
- In the process, the nucleons in the original main-sequence star are converted to iron nuclei.
- In ^{56}Fe, the 26 protons and 30 neutrons are highly ordered compared with 56 free nucleons in the original star, because they are constrained to move together as part of the iron nucleus.
- Thus, the core of the star becomes *more ordered* compared with the original star as the nuclear fuel is consumed.
- The entire universe tends to *greater disorder,* as required by the second law of thermodynamics, because the star radiates energy in the form of photons and neutrinos as it builds its ordered core.

20.4.3 Sequence of Events in Core Collapse

When the mass of the Fe core exceeds the Chandrasekhar limit the core collapses gravitationally. The collapse is *accelerated by two factors:*

1. High-energy γ-rays lead to *photodisintegration of iron into α-particles* by reactions like

   $$^{56}\mathrm{Fe} \rightarrow 13\alpha + 4\mathrm{n},$$

   which is highly endothermic, with $Q = -124.4\,\mathrm{MeV}$.

2. As the density and temperature increase, the rate for the *neutronization reaction,*

   $$\mathrm{p}^+ + \mathrm{e}^- \rightarrow \mathrm{n} + \nu,$$

   is greatly enhanced, which removes electrons from the core.

As we will now show, these factors *remove energy and sources of pressure* rapidly from the core, thus destabilizing it.

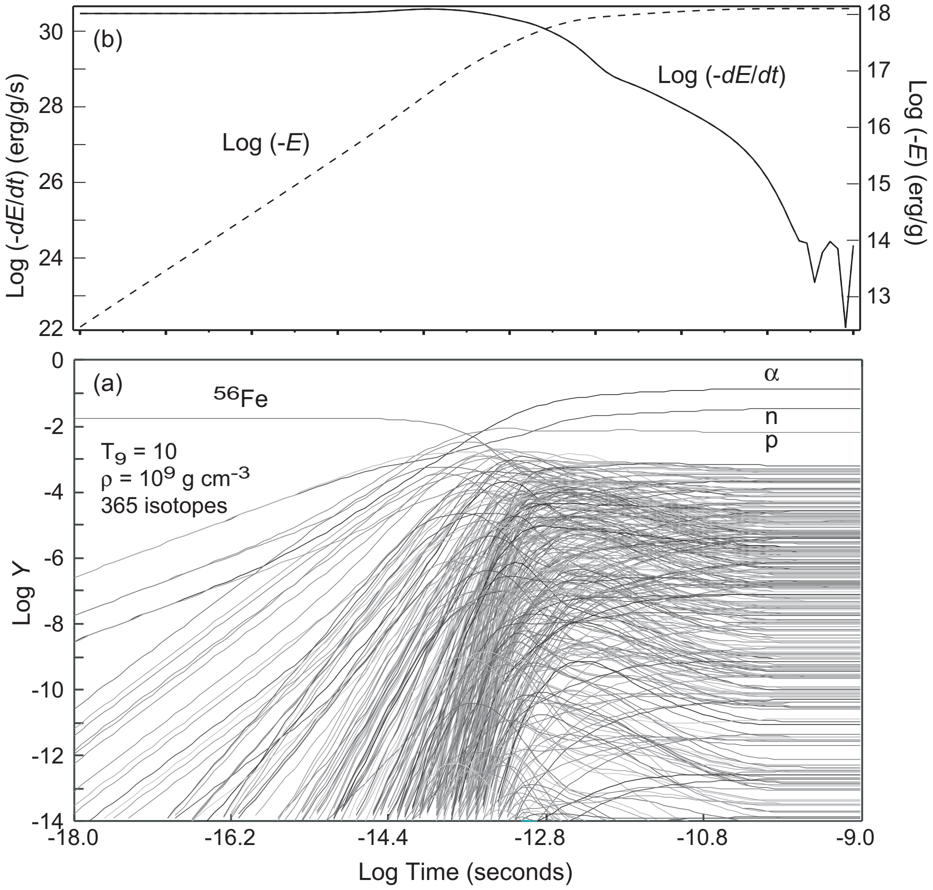
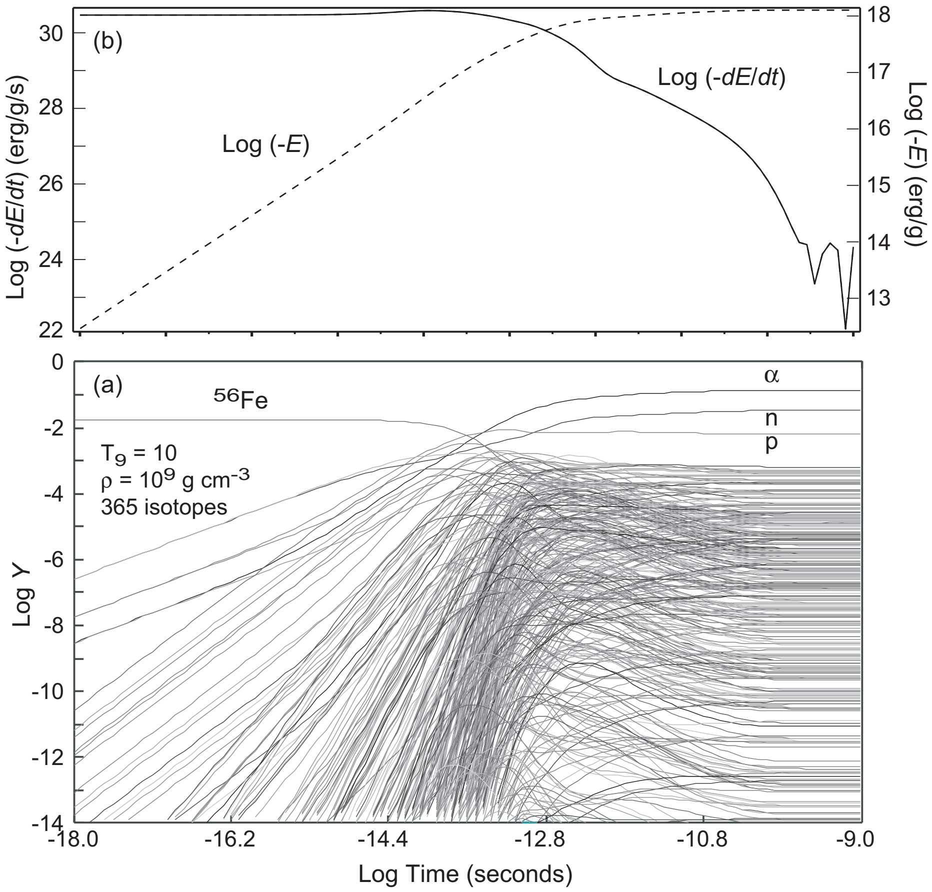


Figure 20.9: Simulated photodisintegration of ^{56}Fe at a temperature of 10^{10} K and density of $10^9\,\mathrm{g\,cm^{-3}}$. (a) Initially only pure ^{56}Fe is present; after $\sim 10^{-12}$ s the original ^{56}Fe has been transformed into 365 isotopes with non-zero populations, but the only species with abundances in excess of 10^{-3} are alpha particles, neutrons, and protons. (b) The rate of energy absorption for this very endothermic reaction.

As the core heats up, high-energy γ-rays are produced rapidly.

- These photodisintegrate iron-peak nuclei, as illustrated in the simulation of Fig. 20.9.
- Hundreds of isotopes are produced .



- The most abundant species at equilibrium under these conditions is α-particles, and
- only α-particles, neutrons, and protons have abundances larger than 10^{-3}.
- Photodisintegration of iron is highly endothermic, as indicated in Fig. 20.9(b).
- For example, $^{56}\mathrm{Fe} \rightarrow 13\alpha + 4\mathrm{n}$ has $Q = -124.4$ MeV.
- This *decreases the kinetic energy of electrons* in the core, which *lowers the pressure* and *hastens the collapse*.

In the collapsing core the *neutronization reaction,*

$$\mathrm{p}^+ + \mathrm{e}^- \to \mathrm{n} + \nu,$$

converts protons and electrons into neutrons and neutrinos.

- This *lowers the pressure* contributed by electrons.
- The *neutrinos escape* the core during the initial collapse
- because their mean free path is much larger than the initial radius of the core.
- These *neutrinos carry off energy*,
  - decreasing core pressure and
  - accelerating collapse even further.
- They also *deplete the lepton fraction* (defined analogous to the electron fraction, but for all leptons).

- The core collapse accelerated by *photodisintegration* and *neutronization*
  - proceeds on a timescale of *tens of milliseconds*,
  - with velocities that are significant fractions of the free-fall velocities.
- The core separates into
  - An *inner core* that collapses subsonically and homologously (Homologous means that the collapse is “self-similar”, in that it can be described by changing a scale factor.)
  - An *outer core* that collapses largely in free-fall, with a velocity exceeding the local velocity of sound in the medium (it is supersonic).
- This collapse is
  - Rapid on timescales characteristic of most stellar evolution,
  - Slow compared with the reaction rates and the core is approximately in equilibrium during all phases of the collapse.

Thus *entropy is constant*, and the highly-ordered iron core before collapse ($S \simeq 1$) remains ordered during the collapse.

- As the collapse proceeds and the temperature and density rise, a point is reached where *because of coherent scattering* the neutrino interactions become sufficiently strong that
  - the *mean free path of the neutrinos becomes less than the core radius*.
  - Shortly thereafter, the time for neutrinos to diffuse outward becomes longer than the characteristic time of the collapse and
  - the *neutrinos are effectively trapped in the collapsing core* (neutrino mean free paths in the collapsed core can become as short as a fraction of a meter).
  - The radius at which this occurs is termed the *trapping radius,* and it is closely related to the neutrinosphere to be defined below.

- Because the collapse proceeds with low entropy,
  - There is little excitation of the nuclei and
  - the nucleons remain in the nuclei until densities are reached where nuclei begin to touch.
- At this point, the collapsing core begins to resemble a gigantic “macroscopic nucleus”.
- This core of extended nuclear matter
  - is a nearly degenerate fermi gas of nucleons, and
  - has a very stiff equation of state because nuclear matter is highly incompressible.
- At this point, the pressure of the nucleons begins to dominate that from the electrons and neutrinos.

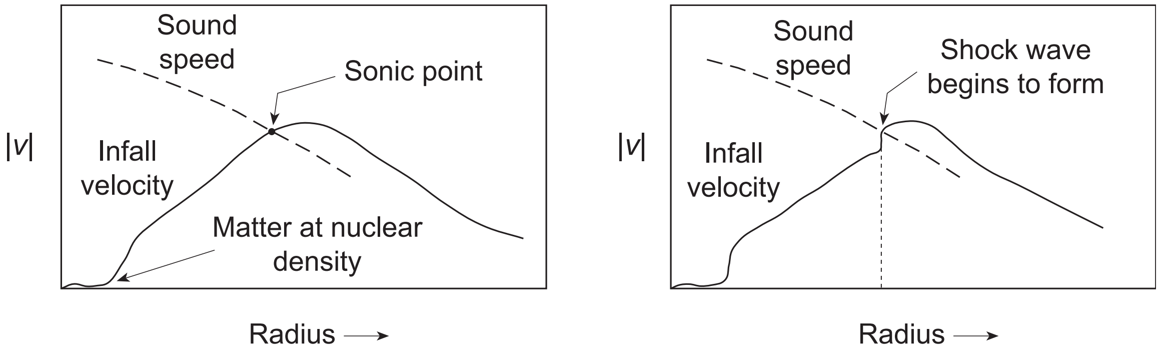


Figure 20.10: The sonic point during collapse (left) and the beginning of shock-wave formation following the bounce (right).

- Somewhat beyond nuclear density,
  - the incompressible core of nearly-degenerate nuclear matter rebounds violently as
  - a pressure wave reflects from the center of the star and proceeds outward.
  - This wave steepens into a shock wave as it moves outward through material of decreasing density (and thus decreasing sound speed),
  - The shock wave forms near the boundary between the subsonic inner core and supersonic outer core (this point is called the *sonic point;* see Fig. 20.10).

In the simplest picture this shock wave would eject the outer layers, resulting in a supernova explosion. This is called the *prompt shock mechanism*.

- The gravitational binding energy of the core at rebound is about 10^{53} erg, and
- the typical observed energy of a supernova (the expanding remnant plus photons) is about 10^{51} erg.
- Thus, only about 1% of the gravitational energy need be released in the form of light and kinetic energy to account for the observed properties of supernovae.

10^{51} erg defines a unit of energy that is termed the *foe:*

$$1\ \text{foe} \equiv 10^{51}\ \text{ergs},$$

with the name “foe” deriving from the first letters of fifty-one ergs.

In more modern usage, this unit is termed the *bethe*,

$$1\ \text{beta} = 1\ \text{foe} \equiv 10^{51}\ \text{ergs},$$

in honor of *Hans Bethe*.

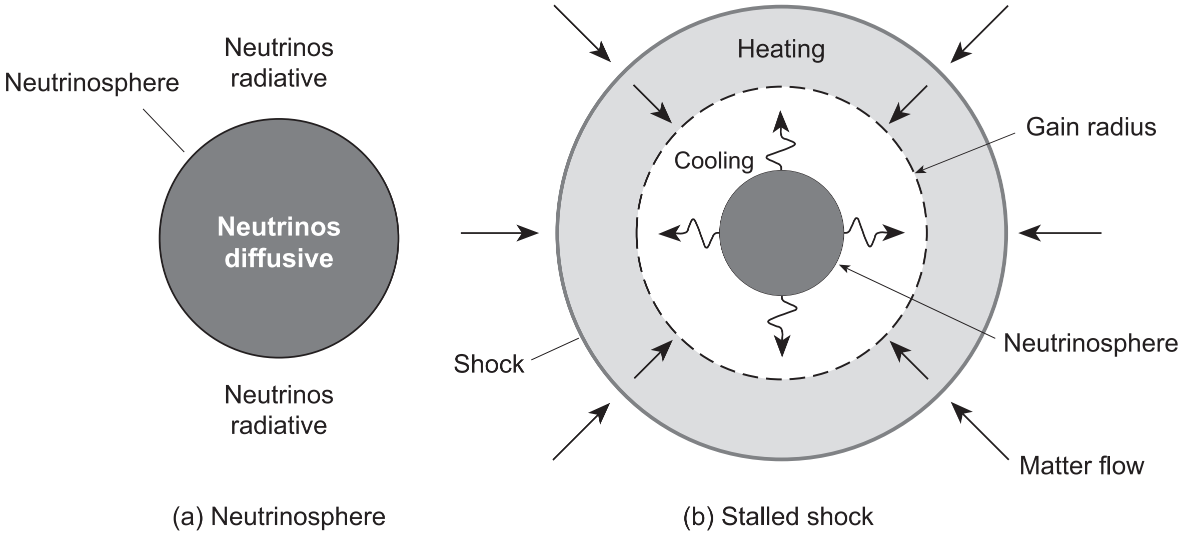


Figure 20.11: (a) Neutrinosphere. (b) Conditions at time of shock stagnation. The neutrinosphere and the gain radius are indicated.

- Unfortunately, a simple prompt-shock mechanism appears to be energetically prohibited:

  1. The shock dissociates nuclei as it passes through the outer core, sapping it of a large amount of energy.
  2. As the shock wave passes into less dense material,

     - the mean free path for the trapped neutrinos increases until the neutrinos can once again be freely radiated from the core.
     - The radius at which neutrinos change from diffusive to radiative is termed the *neutrinosphere*.
     - When the shock penetrates the neutrinosphere a burst of neutrinos is emitted.
     - This carries away large amounts of energy.

This further depletes the shock by lowering pressure behind it.

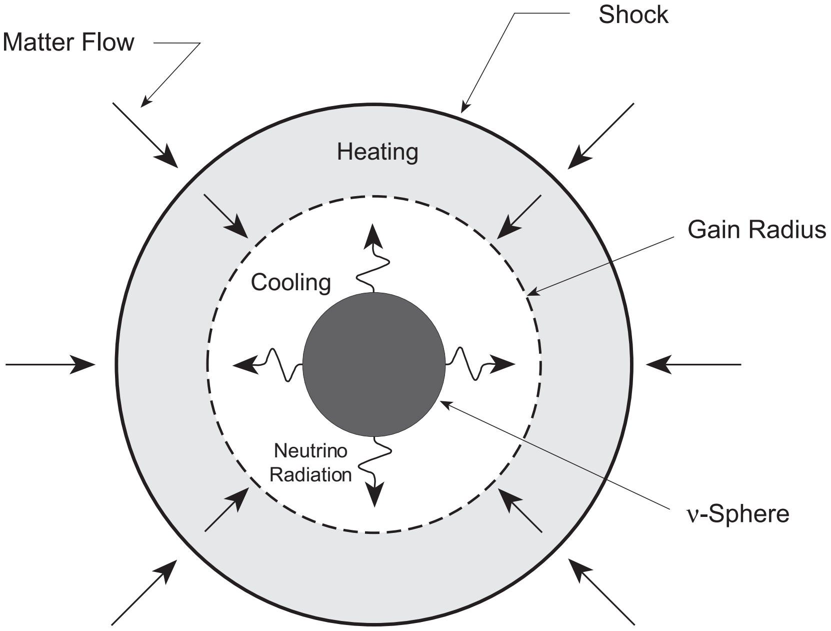


Figure 20.12: Conditions at shock stagnation in a core-collapse supernova. The neutrinosphere and the gain radius are indicated.

- Realistic calculations indicate that the shock wave stalls into an *accretion shock* (a standing shock wave at a constant radius) before it can exit the core, unless the original iron core contains less than about 1.1 solar masses.
  - In a typical calculation, the accretion shock forms at about 200–300 km from the center of the star within about 10 ms of core bounce.
  - Since there is considerable agreement that SN1987A resulted from the collapse of a core having 1.3–1.4 solar masses, the prompt shock mechanism is unlikely to be a generic explanation of Type-II supernovae.

Figure 20.12 illustrates the conditions characteristic of a stalled accretion shock in a core-collapse supernova.

20.4.4 Neutrino Reheating

The idea that neutrinos might play a significant role in supplying energy to eject the outer layers of a star in a supernova event is an old one.

- Failure of the prompt mechanism to yield supernova explosions consistently led to a revival of interest in such mechanisms.
- This evolved into what is generally termed the delayed shock mechanism or *neutrino reheating mechanism*.
- In this picture, the stalled accretion shock is re-energized through heating of matter behind the shock by interactions with neutrinos produced in the region interior to the shock.
- This raises the pressure sufficently to impart an outward velocity to the stalled shock on a timescale of approximately one second, and
- the reborn shock then proceeds through the outer envelope of the star.

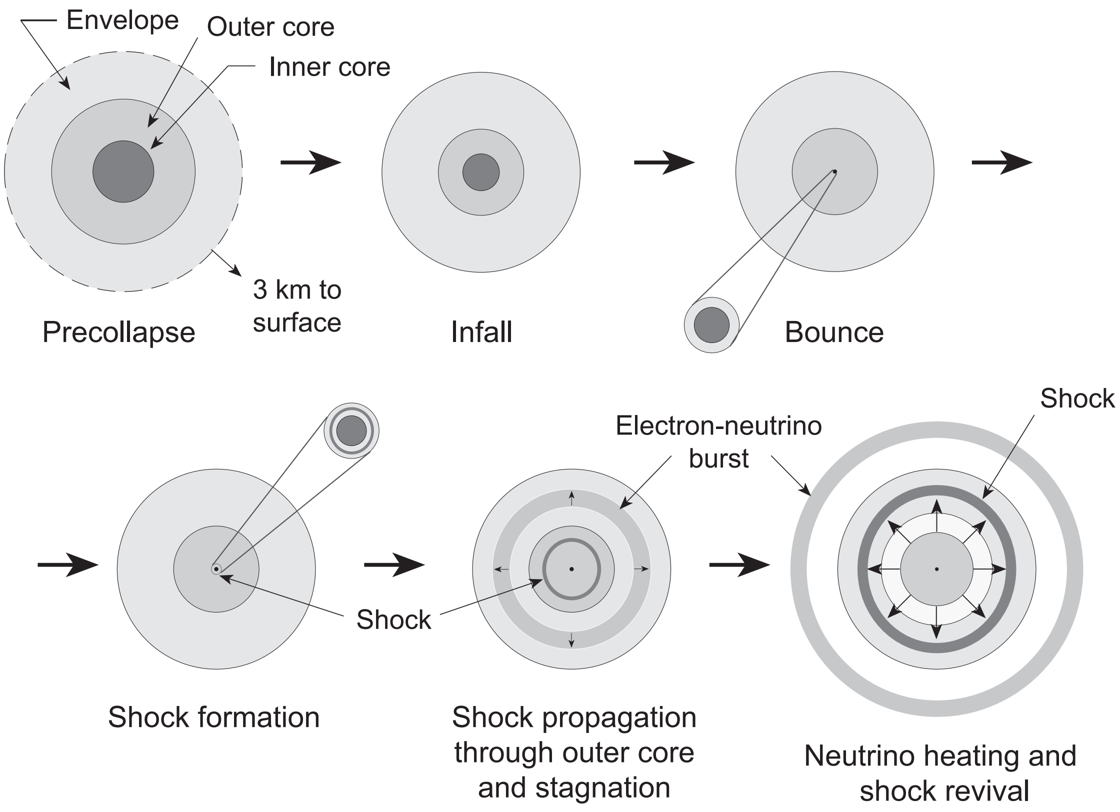


Figure 20.13: Neutrino reheating mechanism for a supernova explosion (after Bruenn). Figures are approximately to scale; the surface of the star would be 3 km from the center on this scale.

The schematic mechanism for the supernova event thus becomes the two-stage process depicted in Fig. 20.13.

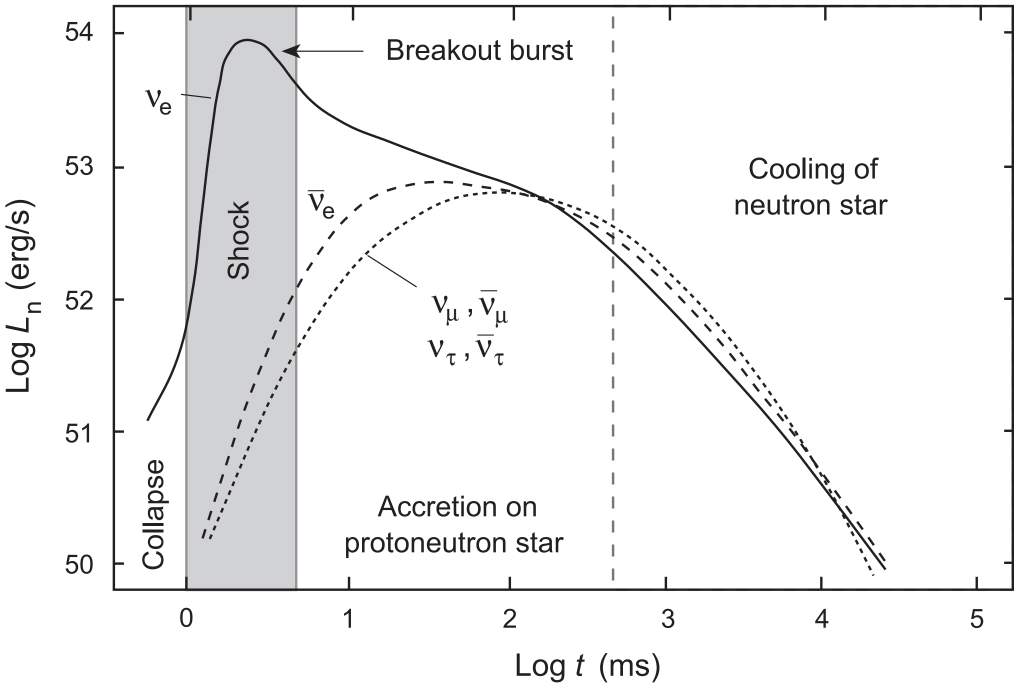
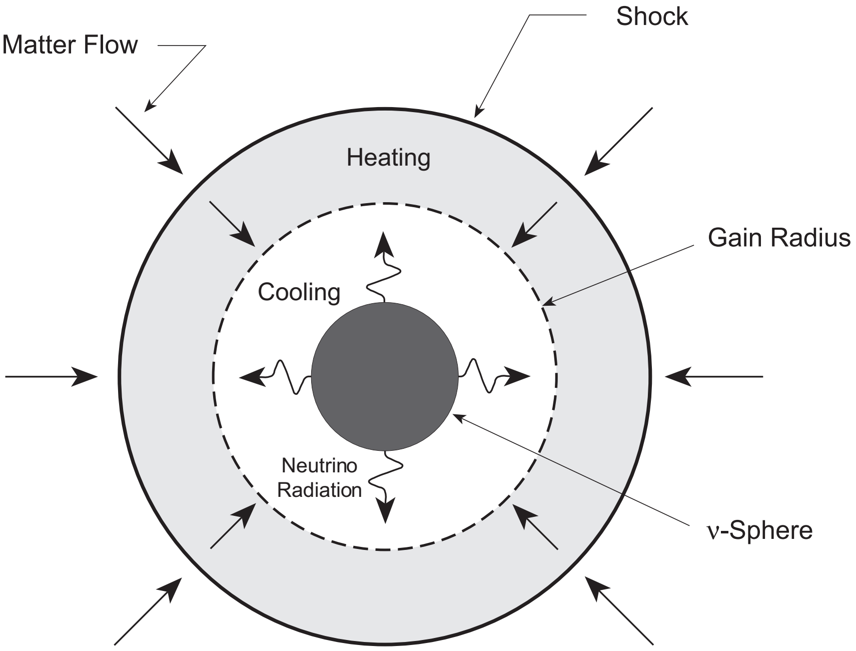


Figure 20.14: Neutrino luminosities in a supernova explosion.

20.4.5 Reheating of Shocked Matter

Neutrinos and antineutrinos are emitted copiously from the hot, dense center of the collapsed core (Fig. 20.14).



Neutrinos can interact with the matter behind the shock. It is useful to introduce two characteristic radii.

- The first we already met: the *neutrinosphere* marks a boundary between diffusive and free-streaming transport.
- The second is associated with the observation that interactions with the matter could either *cool it or heat it*.
  - There is always a radius outside of which the net effect of the neutrino interactions is to heat the matter.
  - This break-even radius is termed the *gain radius*.
  - The *neutrinosphere* and *gain radius* are indicated schematically in the figure above.
  - Shock revival is favored by deposition of neutrino energy between the gain radius and the shock.

20.4.6 Calculations with Neutrino Reheating

The result of a large number of calculations is that

- *Neutrino reheating helps*, but generally does not produce successful explosions without artificial boosts of the neutrino luminosities.
- This suggests that there still are missing ingredients in the supernova mechanism that must be included to obtain a quantitative description.
- One possibility is that *convection interior to the shock* alters the neutrino spectrum and luminosity in a non-negligible fashion.

Let's now turn to a general discussion of *the role that convection might play* in supernova explosions.

20.4.7 Convection and Neutrino Reheating

In Ch. 7 we discussed general conditions under which stars can become *convectively unstable*.

- Substantial convection inside the stalled shock
  - could have significant influence on the possibility of neutrinos reenergizing the shock, and influence
  - quantitative characteristics of a reenergized shock.
- To boost the stalled shock it is necessary for neutrinos to
  - deposit energy *behind the shock front*,
  - but *outside the gain radius*,
- Convective motion inside the shock front could, by *overturning hot and cooler matter*,
  - cause *more neutrino production* and
  - *alter the neutrino spectrum*.
- The convection could also move neutrino-producing matter *outside the neutrinosphere*.
- Then neutrinos that are produced would have a better chance to propagate into the region closely behind the shock where deposition of energy optimizes reheating.

This would provide a method to produce a supernova explosion with the required $\mathscr{O}\left(10^{51}\right)$ ergs of kinetic and visible energy.

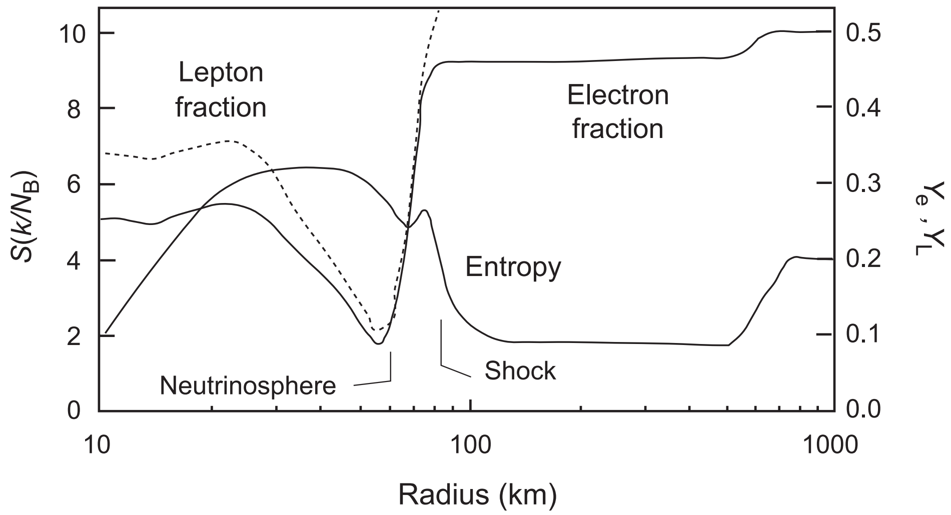


Figure 20.15: Lepton fractions and entropy 6.3 ms after bounce in a supernova calculation. Regions particularly favorable for convective motion are described in the text. The progenitor had a mass of 15 $M_\odot$.

20.4.8 Convectively-Unstable Regions in Supernovae.

Armed with earlier results for predicting *convectively-unstable regions*, let's examine *entropy and electron-fraction gradients* found during shock stagnation in supernova calculations.

- Fig. 20.15 shows results for a 15 $M_\odot$ star 6 ms after bounce. There are *two potentially unstable regions:*
  - A *Schwarzschild unstable region* lying inside the shock front at about 80 km with a *large negative entropy gradient*.
  - A region inside the neutrinosphere where *both the entropy and lepton fraction* exhibit *strong negative gradients* that favor *Ledoux convection*.

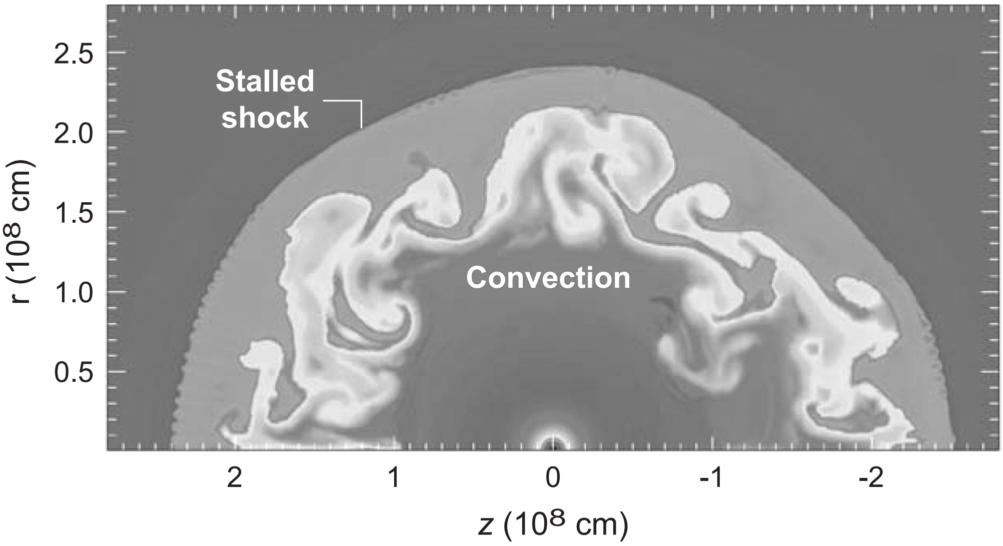


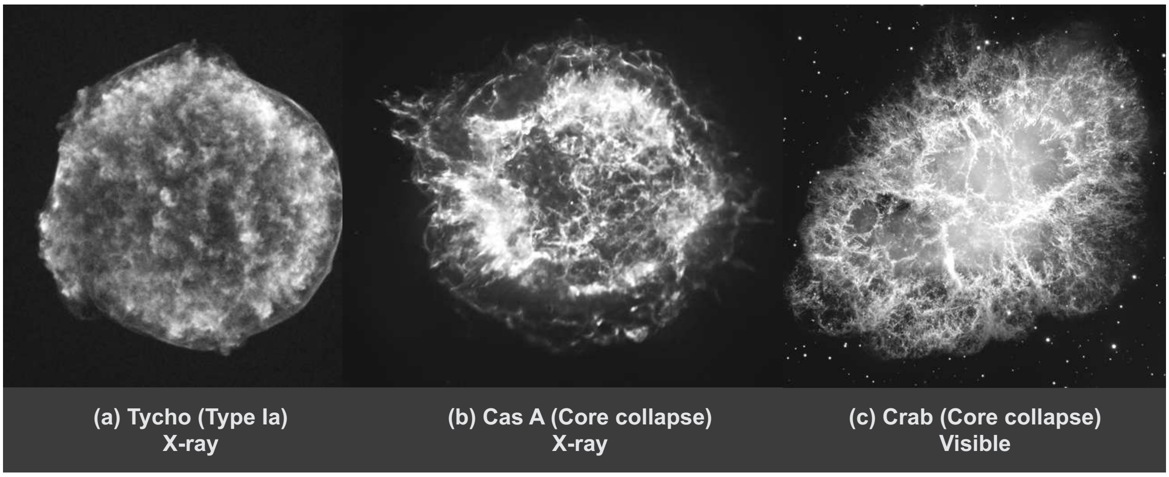
Figure 20.16: 2D core collapse simulation exhibiting violent convection behind a stalled shock. r is the distance from the z axis. Entropy in grayscale, with white maximum and dark gray minimum. The shock is being distorted by the convection beneath it. In modern calculations a *standing accretion shock instability (SASI)* develops associated with deformations of the shock that can be significant in producing successful explosions.

The preceding arguments identify regions that are unstable.

- Whether such regions develop convection, the convective timescale, and the quantitative implications for supernova explosions can only be settled by simulations.
- Many calculations have demonstrated that *convection is significant* in core-collapse supernovae.
- An example is shown in Fig. 20.16, where we see the onset of *spectacular convection below the stalled shock*.
- Such *violent and large-scale convection* can only be modeled using numerical multi-D hydrodynamics simulations.

The present feeling is that

- A completely successful model of core-collapse supernovae will require both
  - Full *3D radiation hydrodynamics* and
  - a full treatment of *neutrino transport*.
- It has not been possible to include both in current codes because of *inadequate computing power*.
- It is thought that the next generation of supercomputers *(exascale computers)* may provide sufficient computational power to determine whether
  - the ingredients described above lead to a succesful model of core-collapse supernovae, or
  - whether they point to the necessity of new physics not yet incorporated into the models.



20.4.9 Remnants of Core Collapse

A core collapse supernova is expected to eject an *expanding supernova remnant*, as illustrated in Figs. (b) and (c) above.

- Unlike a Type Ia explosion, a core collapse supernova is expected to leave behind also a *compact remnant*—either a *neutron star* or a *black hole*.
- Less-massive progenitors lead to neutron stars but
- for more massive stars the end result is a black hole,
  - produced either *immediately*, or
  - with a *time delay* corresponding to accretion on a protoneutron star causing it to collapse to a black hole.

For increasingly-massive progenitors it is expected that less envelope is ejected and more falls back into the black hole.

For masses above about $30\,M_\odot$, current simulations indicate that

- core collapse may lead to *complete fallback* of the outer layers of the star,
- leaving only a black hole with no ejected supernova remnant.
- However, even for direct collapse to a black hole,
    - gravitational waves, and
    - significant neutrino emission

  are expected.

This *direct collapse* to a black hole without a traditional supernova explosion is probably the *general fate of stars more massive than about* $30\,M_\odot$.

Direct collapse to black holes with masses in the $\sim 100 - 250\,M_\odot$ range could be excluded by the *pair instability*, if metallicities are low.

20.4.10 Natal Kicks

Simulations indicate that core collapse explosions are *asymmetric*,

- Hence the compact remnant is expected to receive a *natal kick* in the explosion.
- Neutron stars have been observed with *space velocities as large as* $\sim 1000\ \mathrm{km\,s^{-1}}$, presumably arising from natal kicks in the supernova explosion that produced them.
- For core collapse in more massive stars it is expected that the natal kick is less severe, since less matter is ejected.
- For the collapse of massive cores directly to black holes with no ejected remnant it is often assumed to be zero.

Natal kicks and the amount of ejected matter *affect strongly whether a binary remains bound* if one of the stars undergoes core collapse.

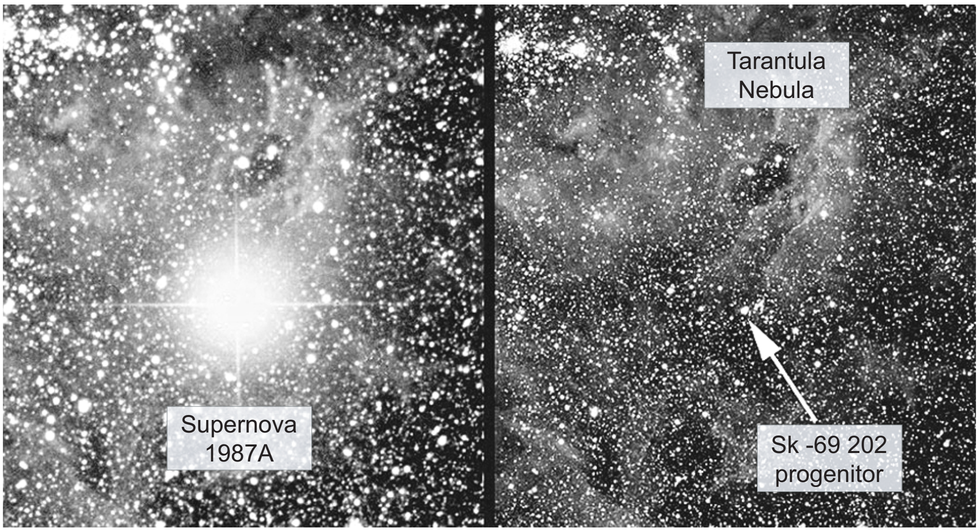
20.5 Supernova 1987A

The *Tarantula Nebula* is a star-forming region in the *Large Magellanic Cloud*, a satellite galaxy of the Milky Way visible in the Southern Hemisphere.

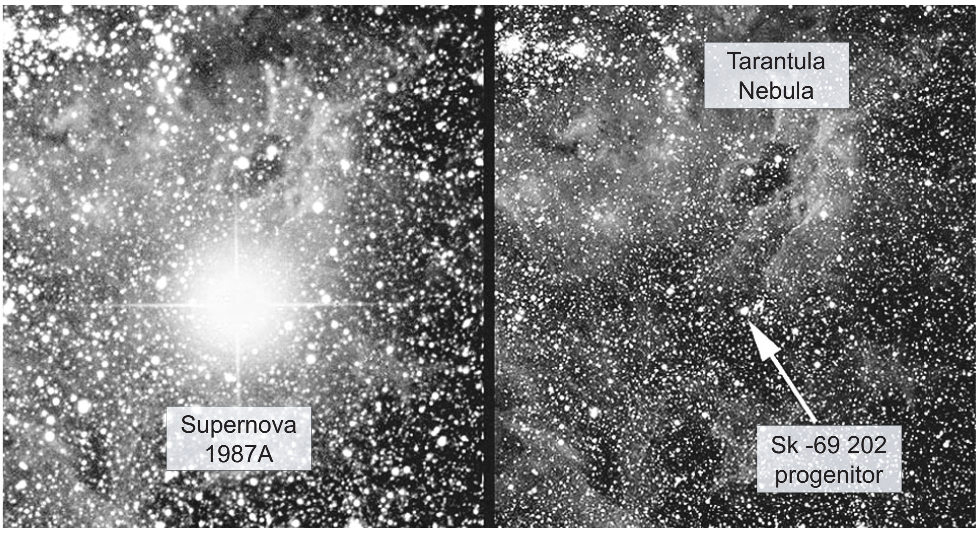
- Some 163,000 years ago the core of a mag 12 blue supergiant in the Tarantula, Sanduleak −69 202, imploded,
  - producing a burst of neutrinos and
  - a shockwave that reached the surface several hours later,
  - sending most of the star's mass hurtling into space and
  - generating a billion-fold increase in luminosity.
- Time passed . . .

On February 23, 1987 on Earth, 163,000 ly away, detectors searching for something else entirely

- saw an unexpected burst of ~ 20 neutrinos.
- A clue to their origin was not long in coming.
- Three hours later, light from the explosion arrived and
- that night observers in Chile and New Zealand were startled to find a "new star", visible to the naked eye, in the Tarantula.
- The progenitor (right) and supernova (left) are shown below.



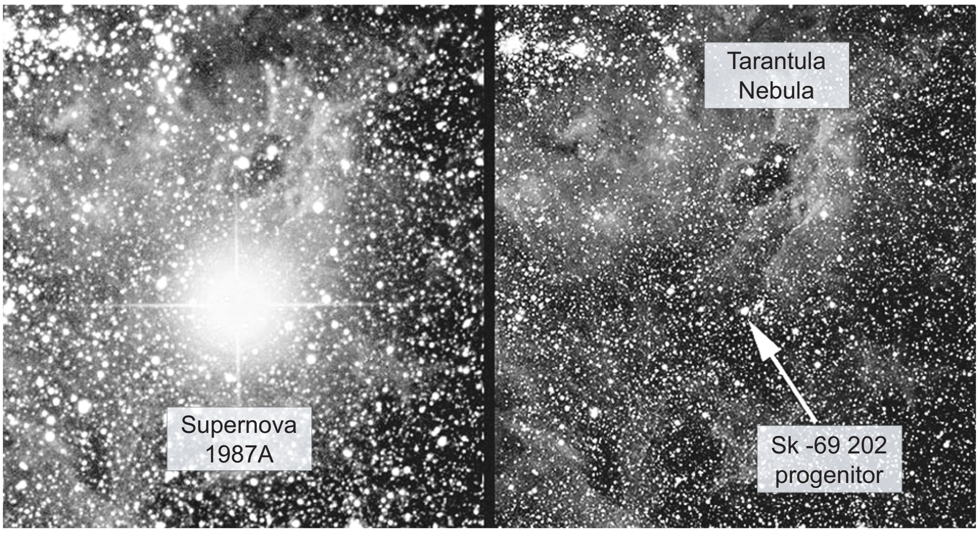
- Thus did neutrinos and light from SN 1987A announce the demise of Sk -69 202 (which could no longer be found after the supernova dimmed) in a core collapse supernova.



The first nearby supernova since the invention of the telescope, SN 1987A has been studied extensively,

- confirming most and
- modifying some

of our understanding of core collapse supernovae.



- This section summarizes how the core collapse mechanism outlined in preceding sections has fared in the light of SN 1987A data.
- In this summary, it is important to remember the distinction made earlier between
  - classification of supernovae with respect to *spectral and lightcurve characteristics*, and
  - classification with respect to *explosion mechanism*.

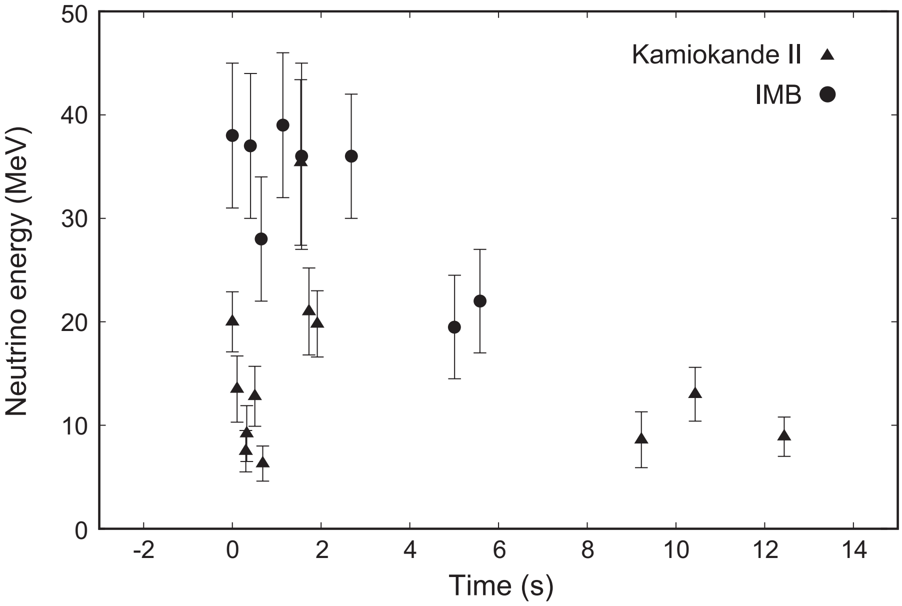
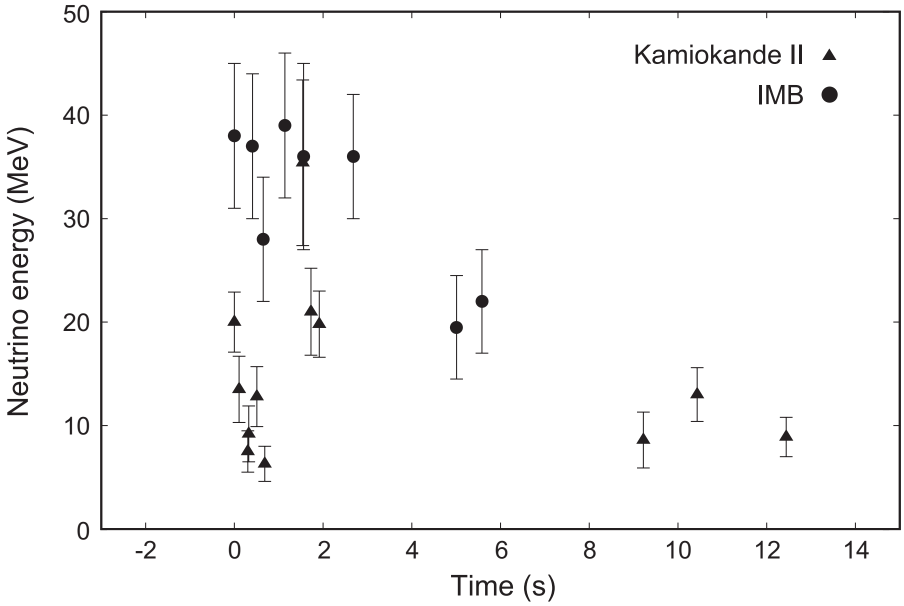


Figure 20.17: Neutrino burst from SN 1987A detected in two water Cerenkov detectors. The inferred direction (with large errors) was consistent with origin of the burst in the Large Magellanic Cloud. This means that the neutrinos passed through the Earth en route to the detectors, which were located in Earth's Northern Hemisphere.

20.5.1 The Neutrino Burst

- Arguably the most important result from SN 1987A was *detection of the neutrino burst*.
- Neutrinos detected in the Kamiokande II and IMB water Cerenkov detectors are shown in Fig. 20.17.



Only 20 neutrinos in total were seen but the general background expected in this plot is very low.

- This low background,
- systematic analysis to rule out the burst being created by a cosmic ray shower, and
- the coincidence of the burst with light from SN1987A (offset by about three hours, as expected)
- leaves no doubt that the neutrinos came from SN 1987A.

Observation of the neutrinos in Fig. 20.17 makes it certain that

- a neutron star or black hole was produced by SN 1987A
- with the release of $\sim 10^{53}$ erg of gravitational energy,

thus *confirming the basic core collapse mechanism*.

Most other observational characteristics of SN 1987A that will now be discussed

- are related to the properties and evolution of the envelope of the progenitor star and are
- only indirectly connected to the collapse of the core defining the explosion mechanism.

Thus they affect observational characteristics, but not the veracity of the basic core collapse mechanism.

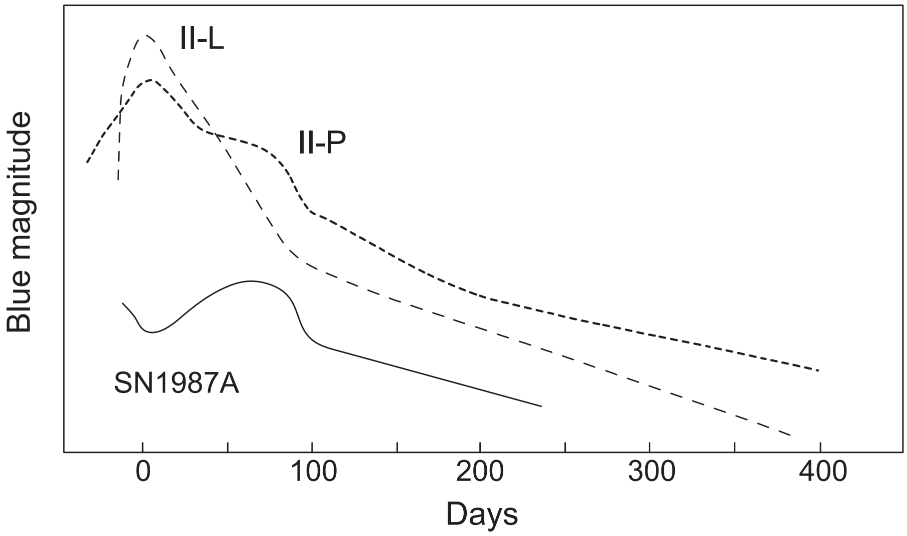


Figure 20.18: Lightcurves for some Type II supernovae.

20.5.2 The Progenitor Was Blue!

The progenitor of Supernova 1987A *came as a surprise* for many because:

- It was widely (but not uniformly) believed at the time that supernova explosions resulted from *core collapse in red supergiant stars*, not blue supergiants like Sk −69 202.
- The early lightcurve of SN 1987A *deviated substantially from that expected for a Type II supernova* (see figure above). For example,
  - the luminosity did not peak until 80 days after the explosion, and
  - SN 1987A was $\sim$ 100 times less luminous than a typical Type II supernova.

Theoretical efforts to understand why the progenitor of SN 1987A was blue when the star exploded focused initially on two possibilities:

1. Extensive *mass loss* in prior evolution.
2. Effects due to the *low metallicity* of the Large Magellanic Cloud (LMC).

Subsequent work indicated that the crucial ingredients required to produce a supernova from a blue supergiant with properties similar to SN 1987A were

- low metallicity,
- a progenitor mass not much greater than $\sim 20\,M_\odot$,
- mass loss of no more than a few solar masses in prior evolution, and
- a tuned prescription for convection.

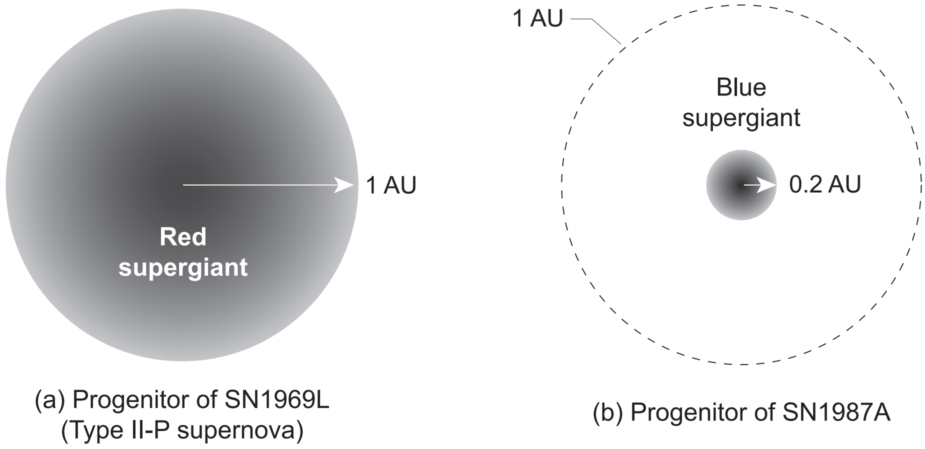
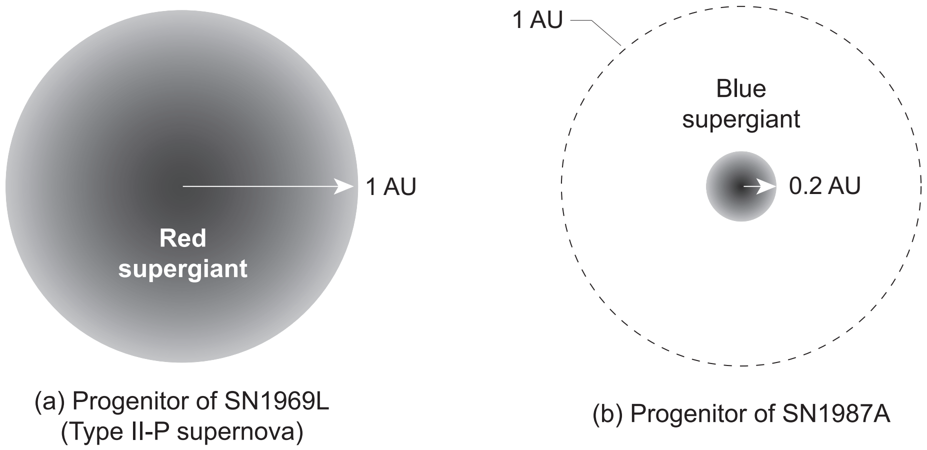


Figure 20.19: (a) Size of a typical red supergiant supernova progenitor; (b) size of the blue supergiant progenitor of Supernova 1987A.

Because Sk −69 202 exploded as a blue supergiant

- the radius of the envelope was much smaller than for a supernova in a red supergiant, as illustrated in Fig. 20.19.
  - A $20 M_\odot$ red supergiant has a radius comparable to the size of the Earth's orbit, but
  - the radius of Sk −69 202 when its core collapsed was only about 20% of the radius of Earth's orbit.
- Thus there was much less envelope for the shockwave to plow through, and
- delay between emission of the initial neutrino burst and the sudden increase in light output when the shock reached the surface was only about 3 hours for SN 1987A.
- For core collapse in a red supergiant of similar mass the time for the shock to reach the surface is likely several times larger than that.



(a) Progenitor of SN1969L (Type II-P supernova)

(b) Progenitor of SN1987A

The much more compact nature of the blue supergiant Sk −69 202 relative to a red supergiant

- also explains the abnormally low luminosity of SN 1987A relative to other Type-II supernovae.
- The primary energy budget may be divided into
  1. production of neutrinos,
  2. ejection of the envelope by the shock, and
  3. powering the lightcurve.
- Neutrino emission dominates the energy budget but it is a property of the core, so it is similar in the two cases.
- For a compact blue supergiant the envelope lies in a deeper gravitational potential than for a red supergiant.
- Therefore, more energy must be expended to eject it, leaving less energy for the subsequent light emission.
- Thus supernovae exploding as blue supergiants are much less luminous than those exploding as red supergiants.

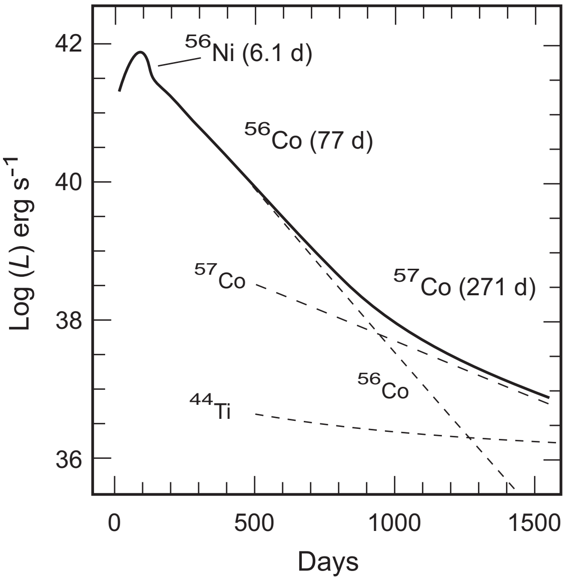


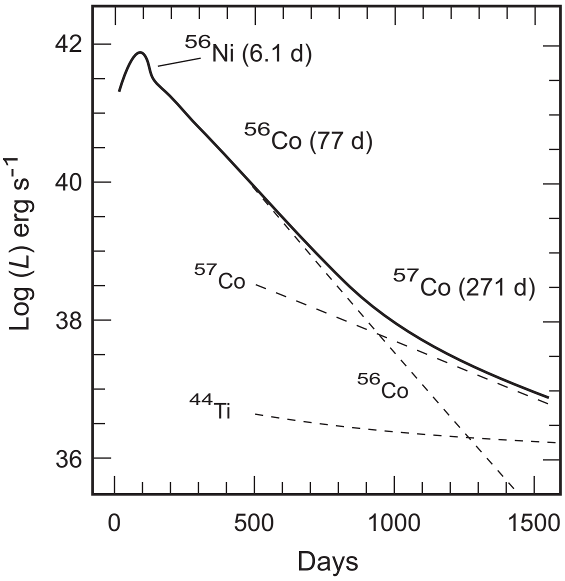
Figure 20.20: Lightcurve of Supernova 1987A (solid curve). The dominant radioactive decays powering the lightcurve at different times are indicated above the lightcurve. The rate of decay for the isotopes powering the lightcurve are indicated by the dashed curves.

20.5.3 Radioactive Decay and the Lightcurve

Initially the lightcurve is powered by the shockwave.

- However, at later times it derives its energy from *radioactive decay of isotopes* produced in the explosion.
- Figure 20.20 illustrates for SN 1987A.
- The lightcurve is for optical photons.
- However, the *energy causing the optical emission* at later times is *supplied primarily by radioactive decay*.

From the shape and height of the lightcurve the *isotopes produced and their abundances* may be inferred.



The initial part of the lightcurve for SN 1987A is accounted for

- if the explosion produced $0.075\,M_\odot$ of ^{56}Ni, decaying by

$$^{56}\text{Ni} \to {}^{56}\text{Co} + e^+ + \nu_e + \gamma$$

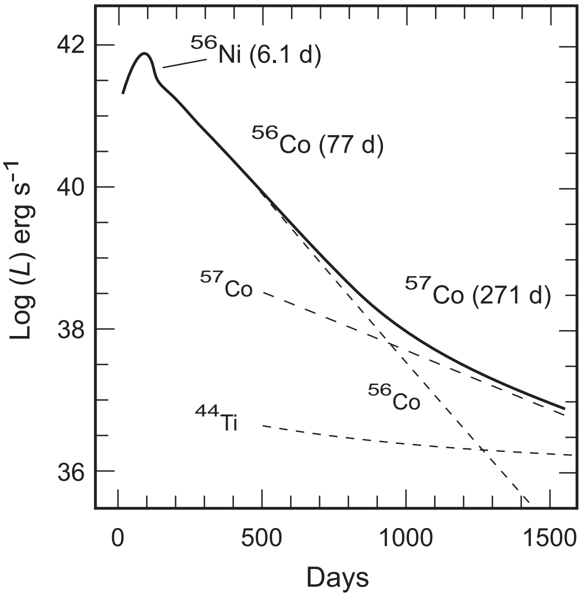
with a 6.1 d halflife.

- Initially optical depth was high and energy released in ^{56}Ni decay produced the *early bump* in the lightcurve.
- Soon after peak luminosity the lightcurve becomes dominated by decay of the ^{56}Co daughter of ^{56}Ni through

$$^{56}\text{Co} \to {}^{56}\text{Fe} + e^+ + \nu_e + \gamma,$$

which has a halflife of 77 d.

- Rate of light production soon becomes determined by rate of energy production and lightcurve slope is then determined by *halflife of the radioactive decay powering it*.



The explosion also produced a much smaller amount of radioactive ^{57}Co, which decays with a 271 d halflife.

- After about 1000 days enough ^{56}Co had decayed away that ^{57}Co decay became dominant.
- The lightcurve then assumed the shallower slope determined by the halflife of ^{57}Co.
- In 2017, thirty years after the explosion, the ^{56}Ni, ^{56}Co, and ^{57}Co had all decayed away.
- The lightcurve was being powered in 2017 by decay of the ^{44}Ti produced in the explosion, which has a 47-year halflife.

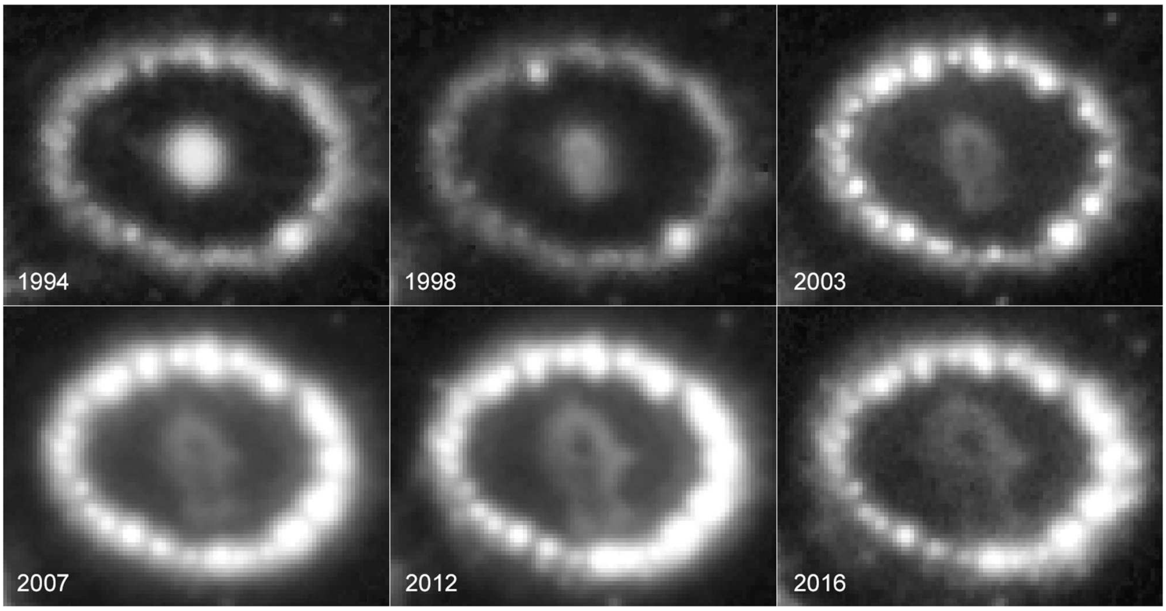


Figure 20.21: Images from 1994 to 2016 showing the collision of the SN 1987A shockwave with a ring of matter emitted by the progenitor before the supernova explosion. See also Fig. 20.22.

20.5.4 Evolution of the Supernova Remnant

Evolution of the expanding remnant of SN 1987A has been studied extensively at multiple wavelengths.

- Fig. 20.21 displays a time lapse of Hubble Space Telescope images from 1994 to 2016.
- The time lapse shows the *collision of the SN 1987A blast wave with a ring of matter* emitted by the progenitor before the supernova explosion.

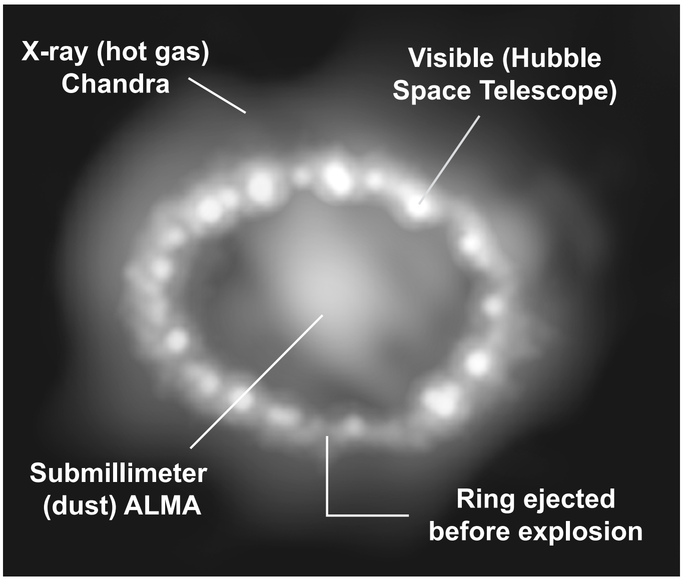
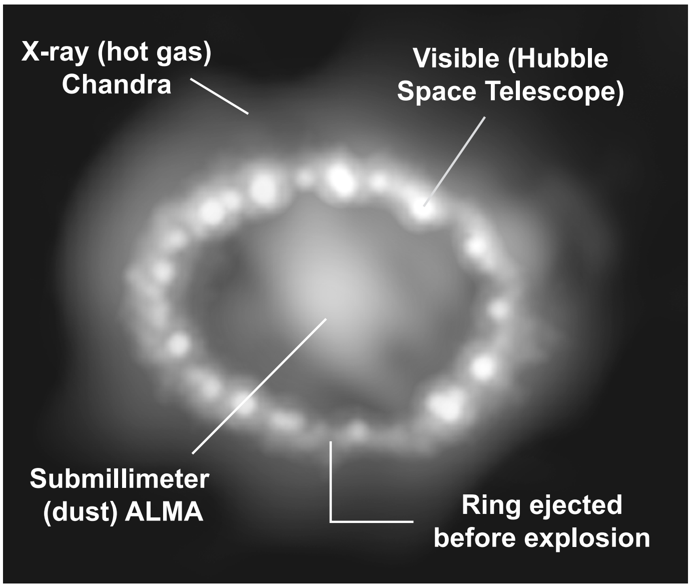


Figure 20.22: Multiwavelength composite 30 years after SN 1987A. In the center dust forming in the supernova remnant is imaged at submillimeter wavelengths by ALMA. Locus of a ring of gas emitted by the star before the explosion is indicated. Brightest clumps in the ring indicate visible light captured by the Hubble Space Telescope that was emitted from the collision of the supernova shockwave with the ring. The more diffuse glow concentrated outside the ring is X-rays imaged by the Chandra X-ray Observatory.

Figure 20.22 shows a multiwavelength composite from 2017.

- In the center dust forming in the supernova remnant is imaged at submillimeter wavelengths by ALMA.
- The locus of a ring of matter ~ 1 ly in diameter that was *emitted by the star before the explosion* is indicated.
- This ring was likely produced by a *late wind from the pre-supernova star* (at least 20,000 years before the explosion)
- A *flash of* UV from the supernova ionized the ring and it has glowed for decades due to electron recombination.



Beginning in the early 2000s the ring began to brighten further as the shockwave from the explosion reached it.

- In the figure above the brightest regions indicate visible light emitted from this collision and captured by the Hubble Space Telescope.
- The more diffuse glow corresponds to X-rays emitted from hot gas produced in the collision and imaged by the Chandra X-ray Observatory.

Concentration of X-rays outside the ring suggests that the shockwave has now passed through the ring and into the less dense matter beyond.

20.5.5 Where is the Neutron Star?

A significant mystery concerns the compact remnant.

- The observed *burst of neutrinos* is a sure sign that a *neutron star or black hole was formed*,
- since *gravitational collapse to a compact remnant* is the only plausible way to release the energy to make the neutrinos.
- From stellar systematics it is estimated that Sk -69 202 had a mass of about $18 M_\odot$ when its core collapsed.
- Simulations indicate that for a progenitor of that mass *the compact remnant should be a neutron star*.
- However, *no clear evidence* for a neutron star has been found, despite extensive searches.

Various explanations have been proposed, none supported conclusively by data. The most plausible are that

- the neutron star is *obscured by dust* and not accreting, making it difficult to see, or that
- the compact object formed was a *black hole and not a neutron star*, which would not be visible if it isn't accreting.

The latter explanation would indicate that we do not fully understand when core collapse forms neutron stars and when it forms black holes.

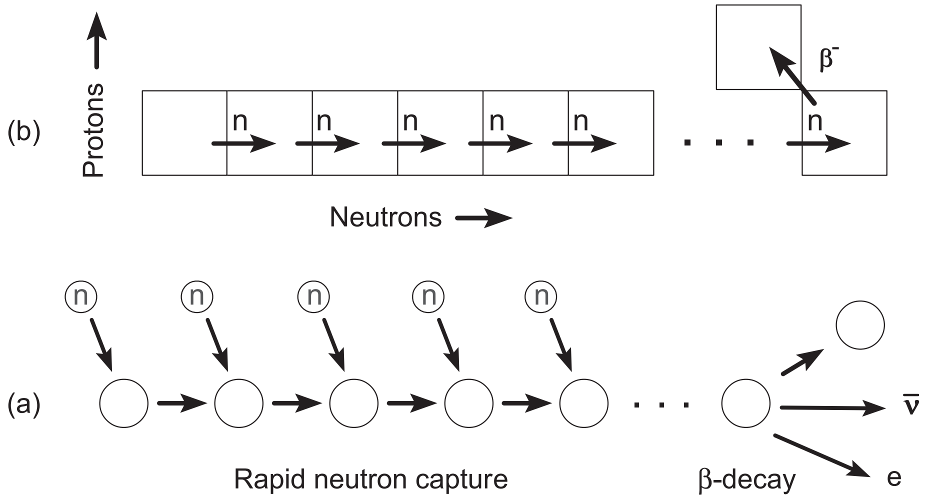


Figure 20.23: Schematic representation of the rapid neutron capture or r-process.

20.6 Producing Heavy Elements: the r-Process

An important question concerns the origin of the heaviest nuclei.

- They *cannot be made by normal charged-particle reactions* in equilibrium in stars because
- the *peak of the binding energy curve* occurs for the *iron-group nuclei* and because of *Coulomb barrier effects*.
- We have noted that *neutron capture reactions* could circumvent the Coulomb barrier problem (since neutrons carry no charge).
- It is thought that many heavy elements are made in the *rapid neutron capture* or *r-process* (Fig. 20.23).

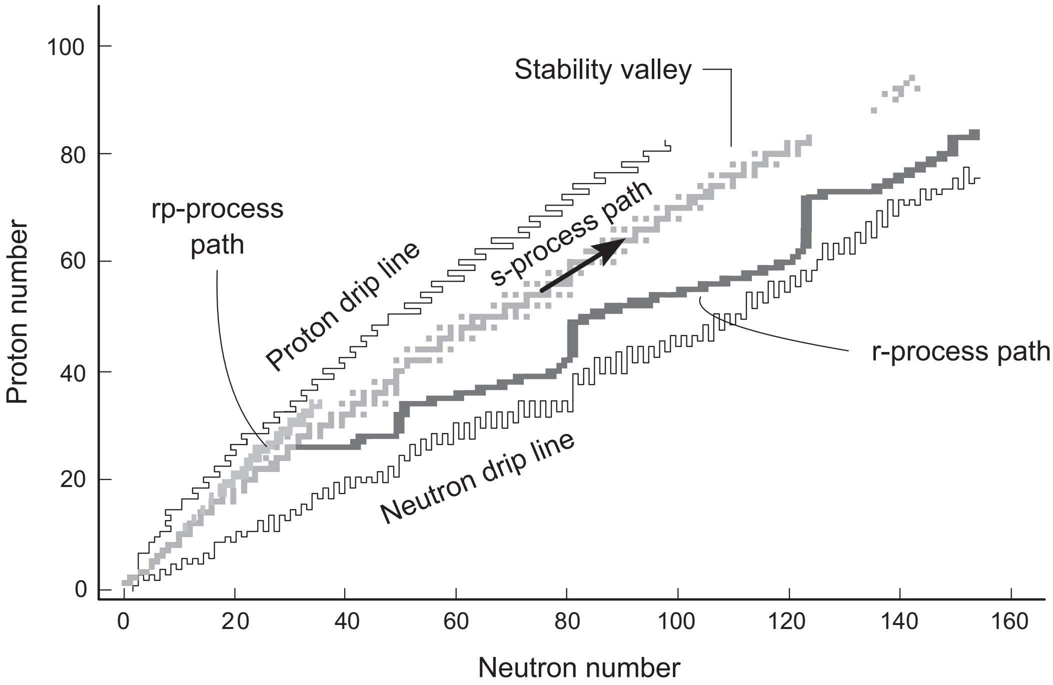


Figure 20.24: The path for the r-process.

This is similar to the s-process in red giants, except that

- Now conditions are such that there is a *very high flux of neutrons* and they can be *captured very rapidly* compared with the time for β decay.
- This takes the population very far to the neutron-rich side of the chart of the nuclides before it begins to β-decay back toward the stability valley (Fig. 20.24).
- Thus the r-process can make *neutron-rich isotopes out of the stability valley* that can't be reached by the s-process.
- Further, the path illustrated in Fig. 20.24 can populate isotopes beyond the gap in the stability valley found near lead and bismuth, thus *accounting for the actinide isotopes*.

The most likely astrophysics sites for the r-process are

- a core-collapse supernova,
- merging neutron stars, or possibly
- jets produced in mergers or core collapse of rapidly rotating stars.

with the first two being the stronger candidates.

- Supernovae and neutron star mergers imply different timescales for r-process nucleosynthesis.
  - A supernova requires a massive star to evolve to gravitational instability of its core, which occurs essentially instantaneously on cosmic timescales.
  - A merger requires a neutron star binary to form by two successive supernova explosions in a massive binary, or by capture of one neutron star by another, and
  - the binary must then spiral together by emission of gravitational waves on a much longer timescale.

> Thus, the r-process associated with mergers has an *inherent time delay.*

The merger delay timescale depends strongly on initial conditions for formation of the binary and in general can be billions of years.

- However, it has been argued that there is a population of fast-merger binaries that can merge on timescales of 10^8 yr or less.
- An open question then is whether binary mergers can account for r-process nuclides observed in low-metallicity stars, which likely formed early in galactic history.

One theme for understanding the origin of r-process nuclei is to ask whether observations suggest that they

- were *produced in a few rare events* (neutron star mergers are relatively rare, occurring maybe only once every million years in a large galaxy),
- or instead were *produced in many more-common events* (core collapse supernovae are much more common, occurring about once every 100 years in a large galaxy).

Some evidence had been accumulating that at least some r-process nuclei were produced in rare events (presumably neutron star mergers).

- The neutron star merger leading to gravitational wave GW170817 and associated gamma-ray burst to be described later gives direct evidence for the *production of large r-process abundances in a single rare event*.
- This has led to much speculation that *neutron star mergers are the primary site of the r-process*.
- However, there are open questions about the rate of neutron star mergers and
- whether they can account for all r-process observations because of the time-delay issues discussed above.

Thus, even if mergers turn are the primary r-process site, it seems likely that core-collapse supernovae contribute to some r-process abundances.

Chapter 21

Gamma-Ray Bursts

The atmosphere absorbs high-frequency photons strongly.

- Thus systematic observation of gamma-rays from space had to await orbiting observatories.
- Because gamma-rays are energetic, they can be produced only in rather unusual and often violent events.
- Therefore, the realization beginning in the 1960s that gamma-rays could be seen from many sources in the sky was a revelation.
- These observations suggested that our Universe was much less sedate and orderly than had often been assumed.
- The most mysterious of the gamma-ray sources were *gamma-ray bursts.*
- These were first observed in the 1960s, but began to be understood only in the 1990s.

As will be discussed in this chapter, it is now believed that gamma-ray bursts represent

- the violent death of a certain class of massive stars, or
- the nearly as violent demise of merging neutron stars.

As such, they are an important part of the story of late stellar evolution, in addition to being of high intrinsic interest because

- they are among the most energetic events that occur in the Universe, and
- because they are potential sources of gravitational waves and perhaps heavy-element synthesis.

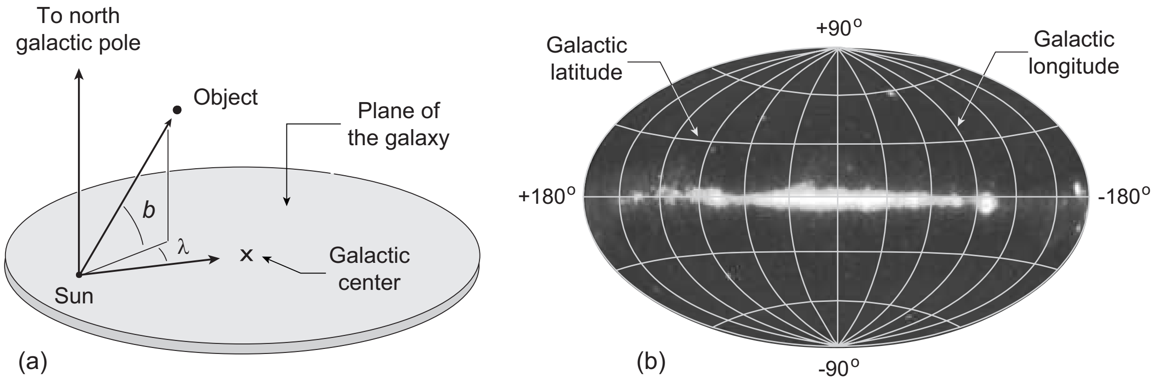


Figure 21.1: (a) Galactic coordinate system. The angle b is the galactic latitude and the angle λ is the galactic longitude, which are related to right ascension and declination by standard spherical trigonometry. (b) The sky at gamma-ray wavelengths in galactic coordinates, with white the most intense and black the least intense. The diffuse horizontal feature at the galactic equator is from gamma-ray sources in the plane of the galaxy.

21.1 The Sky at Gamma-Ray Wavelengths

When seen from space the sky glows in gamma-rays, in addition to the other more familiar wavelengths.

- Figure 21.1 shows the *continuous glow of the gamma-ray sky*, as measured from orbit by the *Compton Gamma-Ray Observatory (CGRO)*.
- In addition to the steady gamma-ray flux illustrated in Fig. 21.1, *sudden bursts*, which can be
  - as short as tens of milliseconds and
  - as long as several minutes,

  are observed.

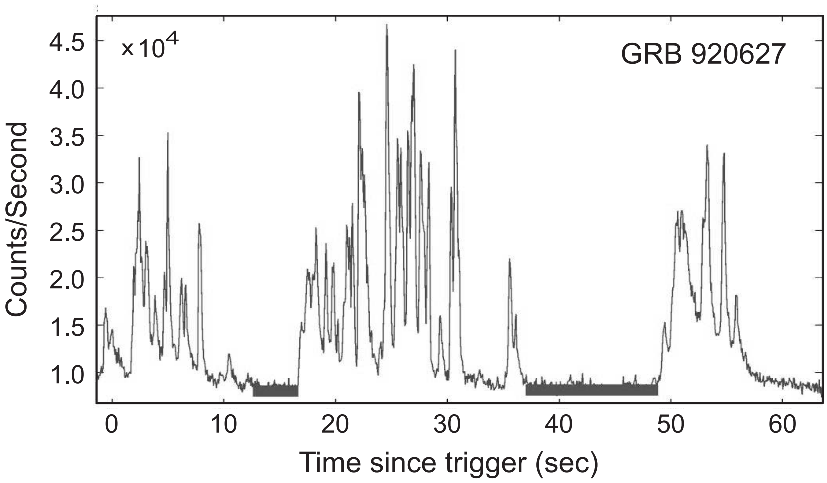


Figure 21.2: Time profile of a gamma-ray burst.

Figure 21.2 displays the time profile for a typical burst event.

- These gamma-ray bursts were discovered unexpectedly in the 1960s by gamma-ray detectors aboard satellites.
- These satellites were testing the feasibility of detecting gamma-rays from nuclear explosions violating test bans treaties.
- Quite surprisingly, the satellites began to see strong bursts of gamma-rays coming from *above*.
- These gamma-ray bursts (GRBs) were for several decades a great puzzle.,

As will now be discussed, newer observations have led to a much deeper understanding of these remarkable events.

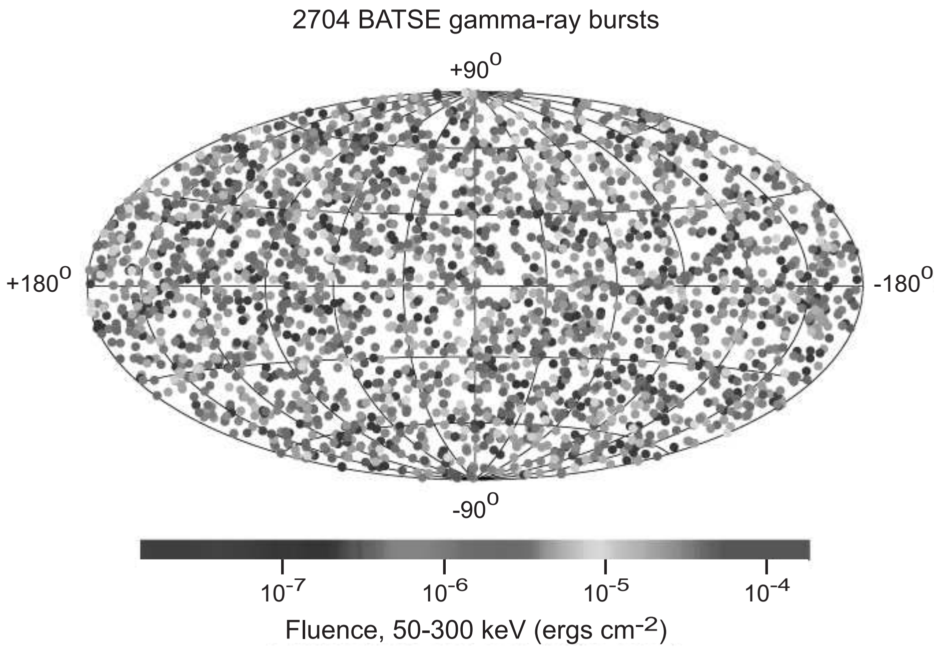


Figure 21.3: Location on the sky of 2704 gamma-ray bursts recorded by the Burst and Transient Source Experiment (BATSE) of the Compton Gamma-Ray Observatory (CGRO), plotted in galactic coordinates with the grayscale indicating the *fluence* (energy received per unit area) of each burst.

About one burst a day is seen by orbiting observatories.

- Figure 21.3 shows the position of 2704 gamma-ray bursts observed by the CGRO.
- The distribution of GRB events is *highly isotropic over a broad range of fluences* (energy received per unit area).
- This argues strongly that they occur at *cosmological distances*—hundreds of megaparsecs or greater).
- The origin of gamma-ray bursts far outside out galaxy will be confirmed more directly below.

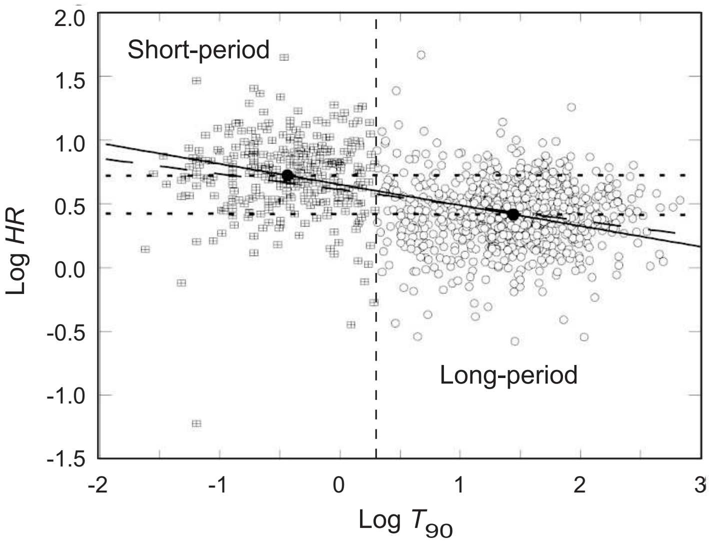


Figure 21.4: Hardness *HR* (a parameter measuring the propensity to contain higher-energy photons) of the spectrum vs. burst duration, illustrating separation of the GRB population into long, soft bursts and short, hard bursts. T_{90} is the time from burst trigger for 90% of the energy to be collected.

Figure 21.4 illustrates two classes of gamma-ray bursts:

1. *Short-period bursts,* which
   - last less than two seconds and
   - exhibit harder (higher-energy) spectra.
2. *Long-period bursts,* which
   - typically last from several seconds up to several hundred seconds and
   - have softer (lower-energy) spectra.

These two classes share many common features but *their differences suggest* that they arise from *two different mechanisms*.

21.2 Localization of Gamma-Ray Bursts

The first step in understanding what causes gamma-ray bursts was to *pin down the astrophysical environment* in which they originate.

- Could they be associated with known galaxies or with specific events like supernova explosions, for example?
- BATSE observations in the 1990s had *angular resolutions no better than several degrees*.
- Thus it was difficult to know exactly where to point telescopes to find evidence associated with the gamma-ray burst at other wavelengths.

Help in this regard came from a small satellite looking not at gamma-rays, but at X-rays.

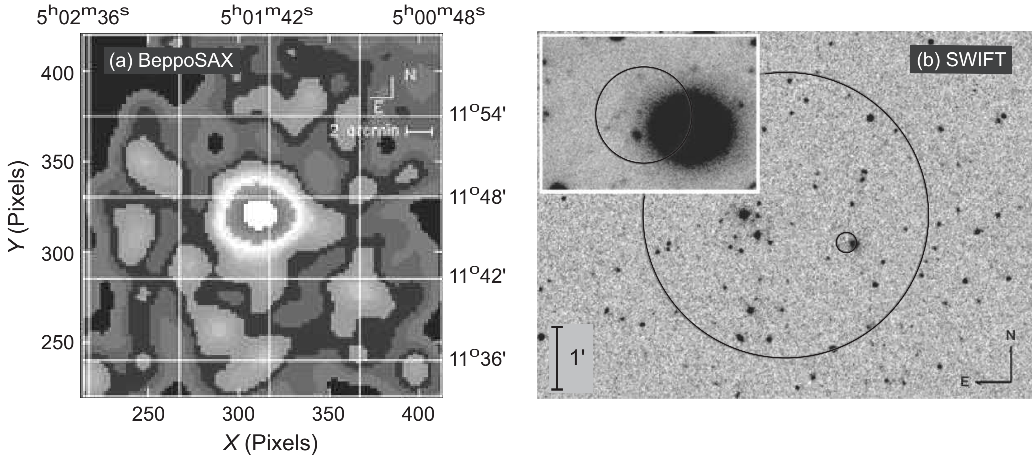
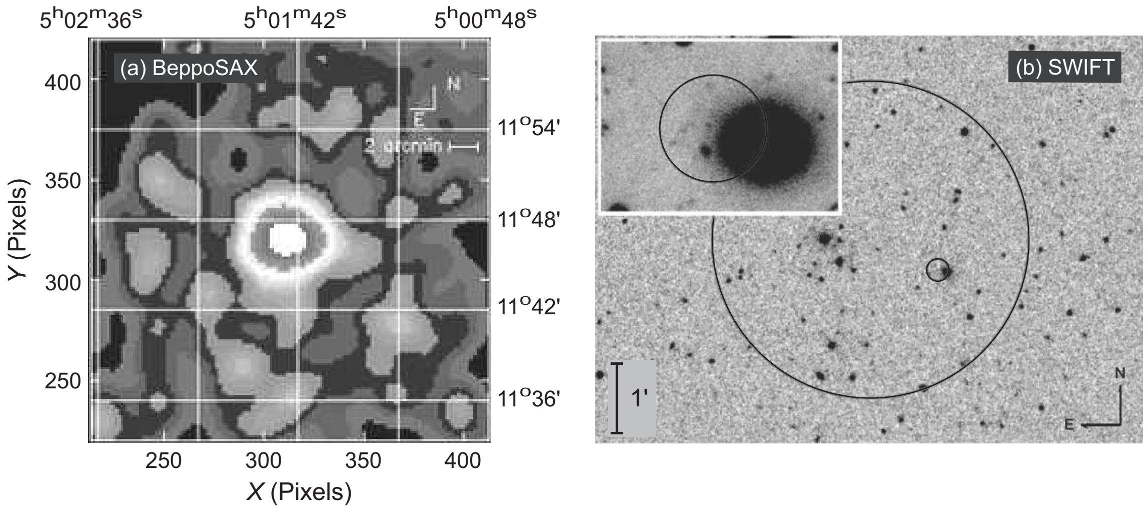


Figure 21.5: (a) First localization of an X-ray afterglow for a GRB by the BeppoSAX. (b) Optical association of short-period GRB 050509B with an elliptical galaxy at $z = 0.225$ by SWIFT. The larger circle is the error circle for the Burst Alert Telescope (BAT). The smaller circle is the error circle for the X-Ray Telescope (XRT), which was slewed to point at the event when alerted by the BAT. The XRT error circle is shown enlarged in the inset at the upper left, suggesting that the GRB occurred on the outskirts of the large elliptical galaxy (dark oval) partially overlapped by the XRT error circle.

In the late 1990s it became possible to correlate some GRBs with other visible, RF, IR, UV, and X-ray sources.

- This was enabled initially by a satellite called BeppoSAX that could localize X-ray transients after a GRB with 2 arc-minute resolution, within hours.
- This permitted other instruments to look quickly at the burst site at multiple wavelengths.
- For the first time transient sources (*“afterglows”*) at other wavelengths could be correlated with a burst.
- Figure 21.5(a) shows an X-ray transient observed by BeppoSAX after a long-period gamma-ray burst.



- Figure (b) above shows a corresponding localization for a short-period burst by the SWIFT satellite (see caption on the previous slide).
- *Redshifted spectral lines* were observed in the transients after the burst.
- For the first time this allowed *reliable distances* to be assigned to gamma-ray bursts.
- These observations showed conclusively that gamma-ray bursts are occurring *at cosmological distances*.
- Thus GRBs must emit *enormous power at gamma-ray wavelengths* to be visible at such distances.
- This raises challenging questions concerning the *source of all that power*.

- Typical stellar spectra are *thermal*, meaning that they arise from a gas in thermal equilibrium and obeying a Planck radiation law.

- To continue our discussion of gamma-ray bursts we must introduce the important idea of a *non-thermal emission spectrum*.

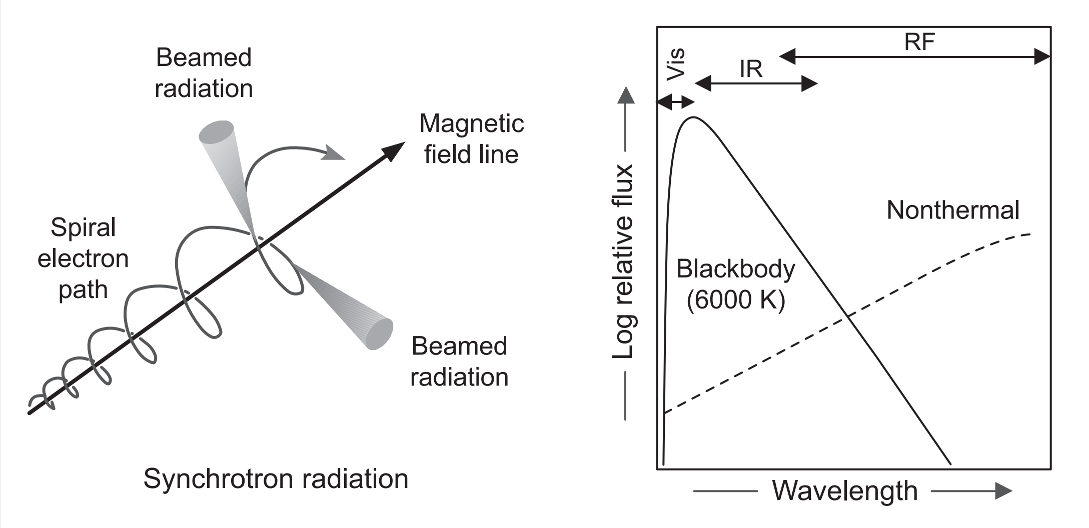


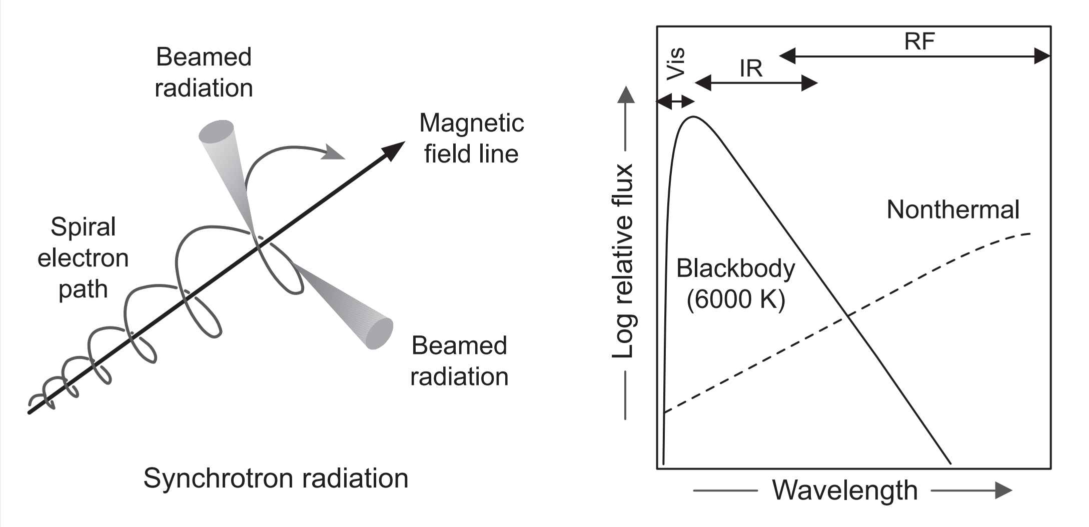
Figure 21.6: Thermal and nonthermal emission

21.2.1 Nonthermal Emission

The Planck law describes *thermal emission,* characterized by emission of radiation from a hot gas in thermal equilibrium; the resulting spectrum is a *blackbody spectrum.*

- Planck law curves for thermal emission peak at some wavelength, and fall off rapidly at longer and shorter wavelengths (curve "Blackbody" in Fig. 21.6).
- The position of the peak moves to shorter wavelength as the gas temperature is increased (the Wien law).
- Light from most stars, and light from normal galaxies, is dominantly thermal in character.

Sometimes *nonthermal* emission is observed, with a spectrum that increases in intensity at very long wavelengths.



The most common form of nonthermal emission in astronomy is *synchrotron radiation.*

- Created when high-velocity electrons (or other charged particles) in strong magnetic fields follow a spiral path around the field lines, radiating their energy in the form of *highly-beamed and polarized light* (figure above left).
- The figure above right contrasts a thermal spectrum characteristic of 6000 K and nonthermal emission.
- The wavelength of the emitted synchrotron radiation is related to how fast the charged particle spirals in the magnetic field.
- Thus, as the particle emits radiation, it *slows and emits longer wavelength radiation*.
- This explains the *broad distribution in wavelength* of synchrotron radiation relative to thermal radiation.

Nonthermal emission is less common than thermal emission in astronomy, but

- The presence of a nonthermal component in a spectrum typically signals *violent processes* and *large accelerations of charged particles*.
- High-frequency synchrotron radiation also implies the presence of *very strong magnetic fields*, since the frequency increases with tighter electron spirals, which are characteristic of strong fields
- The resulting synchrotron radiation has a *nonthermal spectrum and is partially polarized*.
- It is strongly focused in the forward direction by *relativistic beaming*.
- Fluctuations of the jet in time will also be compressed into shorter apparent periods by relativistic effects.
- For an observer in the general direction of a jet, these effects will exaggerate both the apparent intensity and the time variation of the nonthermal emission.
- Thus, the nonthermal part of the continuum emission originates largely in the synchrotron radiation produced in the jets.
- The thermal continuum is typically produced in the accretion disk and surrounding matter that it heats.

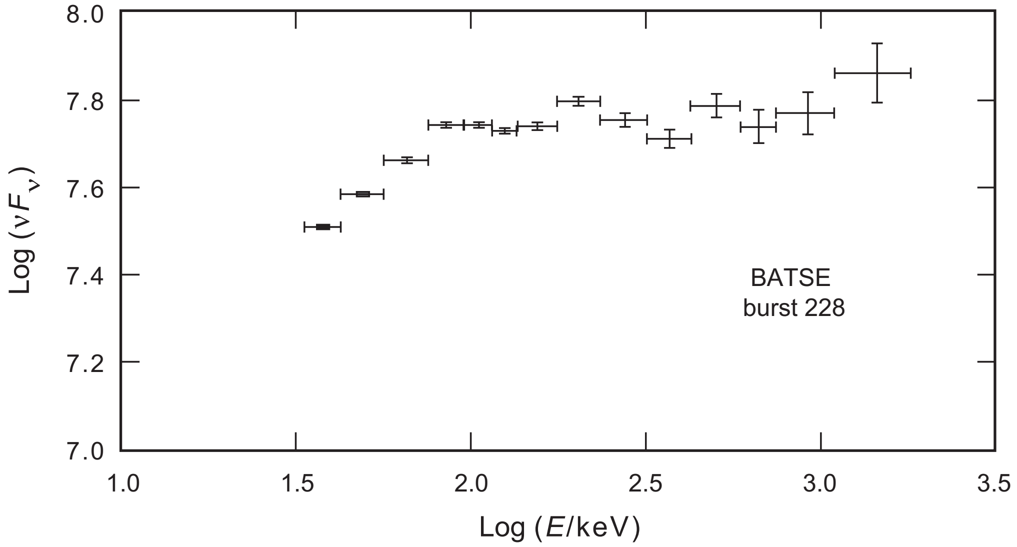


Figure 21.7: Spectrum of a gamma-ray burst. The spectrum is nonthermal.

21.3 Generic Characteristics of Gamma-Ray Bursts

It is now thought that gamma-ray bursts have the following general characteristics:

- *Cosmological origin:* The isotropic distribution of gamma-ray bursts suggests a cosmological origin, which has been confirmed by redshift measurements on emission lines in GRB afterglows.

  Known spectroscopic redshifts for GRBs range up to $z = 8.2$ for GRB 090423.

- *Nonthermal spectrum:* The spectrum is not thermal. Fig. 21.7 illustrates a typical GRB spectrum.

- *Duration and time structure:* The lengths of individual bursts vary from about 0.01 seconds to several hundred seconds, and their time structure can range from smooth to millisecond fluctuations (with the latter implying a compact source).

- *Ultrarelativistic jets:* The gamma rays are strongly beamed, implying emission from tightly-collimated, ultra-relativistic jets. Furthermore, the gamma-rays must suffer little interaction with surrounding matter before escaping, as will be discussed further shortly.

- *Two classes of bursts:* As already noted, there appear to be two classes of bursts: long-period and short-period, with sufficient differences to suggest that they occur through distinct mechanisms.

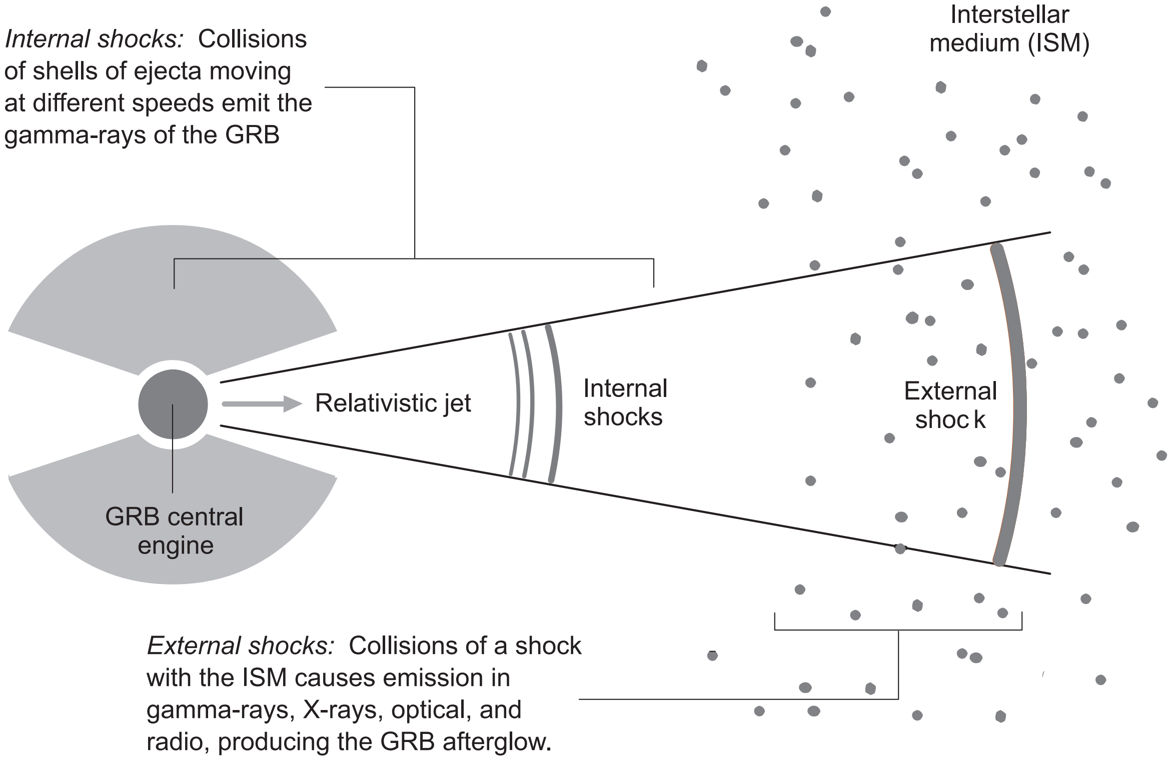

Internal shocks: Collisions of shells of ejecta moving at different speeds emit the gamma-rays of the GRB
Interstellar medium (ISM)
Relativistic jet
Internal shocks
External shoc k
GRB central engine
External shocks: Collisions of a shock with the ISM causes emission in gamma-rays, X-rays, optical, and radio, producing the GRB afterglow.

Figure 21.8: Relativistic fireball model for afterglows following gamma-ray bursts. Internal shocks in the ultrarelativistic jet produce the gamma-rays; the external shocks resulting from the jet impacting the interstellar medium produce the afterglows.

- *Afterglows and fireballs:* The transients (afterglows) observed after gamma-ray bursts
  - can be explained reasonably well by the *relativistic fireball model* illustrated in Fig. 21.8,
  - where deposition of energy by some central engine
  - initiates a *fireball expanding at relativistic velocities* that is responsible for the afterglows.

21.3.1 Necessity of Ultrarelativistic Jets

Gamma-ray bursts *must involve ultrarelativistic jets* because observed prompt emission is *nonthermal*.

- If the jet were not ultrarelativistic the ejecta would be optically thick to pair production for energies less than a few hundred keV, which would thermalize the energy.
- The requirement that gamma-ray bursts be produced by ultrarelativistic jets (and not thermalized photons) can be understood in terms of the opacity of the medium with respect to formation of electron–positron pairs by $\gamma\gamma \rightarrow e^+e^-$.

Let’s elaborate on this crucial point.

21.3.2 Optical Depth for a Nonrelativistic Burst

We first assume that the burst involves nonrelativistic velocities.

- The initial spectrum is *nonthermal*.
- The number of counts $N(E)$ as a function of gamma-ray energy can be approximated (roughly, but sufficient for this estimate) for particular ranges as a power law,

  $$N(E)dE \propto E^{-\alpha}dE,$$

  where the *spectral index* α is approximately equal to 2 for typical cases.
- Because the observed spectrum is nonthermal the medium must be *optically thin* (low opacity),
- since scattering in an optically-thick (that is, highly-opaque) medium would quickly thermalize the photons.

Let's consider the optical depth for pair production associated with a typical gamma-ray burst to see whether this condition can be fulfilled.

For the reaction $\gamma\gamma \rightarrow e^+e^-$ to occur,

- Energy conservation requires the two photons with energies E_1 and E_2, respectively, to satisfy $(E_iE_2)^{1/2} \gtrsim m_e c^2$, where m_e is the electron mass.
- Let f be the fraction of photon pairs that fulfill this condition.
- The optical depth with respect to $\gamma\gamma \rightarrow e^+e^-$ is then

  $$\tau_0 = \frac{f\sigma_T F D^2}{R^2 m_e c^2} \simeq \frac{f\sigma_T F D^2}{\delta t^2 m_e c^4},$$

  where

  - $\sigma_T = 6.652 \times 10^{-25}\,\mathrm{cm}^2$ is the Thomson scattering cross section for electrons,
  - F is the observed fluence for the burst,
  - D is the distance to the source, and
  - R is its size, which can be related to the observed period δt for time structure in the burst by $R = c\delta t$.
- A typical optical depth estimated using this formula is *enormous* ($\tau \sim 10^{14}$).
- Therefore it is completely inconsistent with the $\tau \lesssim 1$ required by the nonthermal GRB spectrum.

Nonrelativistic jets won't do; what about relativistic jets?

21.3.3 Optical Depth for an Ultrarelativistic Burst

The above considerations will be altered in two essential ways if the burst is instead ultrarelativistic with a Lorentz factor $\gamma \gg 1$, so that special-relativistic kinematics apply:

1. The *blueshift* of the emitted radiation will modify the fraction f of photon pairs that have sufficient energy to make electron–positron pairs.
2. The *size* R of the emitting region will be altered by relativistic effects.

Specifically,

- The observed photons of frequency ν and energy $E = h\nu$ have been blueshifted from their energy in the rest frame of the GRB by a factor γ.
- Thus the source energy E_0 was lower than the observed energy E by a factor of γ^{-1} and $E_0 \sim h\nu/\gamma$.
- This means that fewer photon pairs have sufficient energy in the rest frame of the GRB to initiate $\gamma\gamma \to e^+e^-$ than was inferred from the observed energy assuming nonrelativistic kinematics.
- From the spectrum

$$N(E)dE \propto E^{-\alpha}dE,$$

  so f should be multiplied by a factor of $\gamma^{-2\alpha}$.

- Furthermore, relativistic effects increase the size of the emitting region by a factor of γ^2 over that inferred from the time period δt,
- so R should be multiplied by a factor of γ^2.
- Incorporating these corrections, the ultrarelativistic modification of is
$$\tau \simeq \frac{\tau_0}{\gamma^{4+2\alpha}},$$
where τ_0 is the result with no ultrarelativistic correction.
- Thus, even if τ_0 is very large an optically-thin medium can be obtained if γ is large enough.

Typical estimates are that an optically thin medium requires $\gamma \sim 100$ or larger.

Observational confirmation that gamma-ray bursts are associated with the large values of γ deduced from the preceding theoretical analysis comes from

- the observed location of *breaks* in the lightcurves for afterglows.
- These breaks are thought to indicate the time when the initially-relativistic afterglow begins to slow rapidly through interactions with the interstellar medium.
- This in turn can be related to the opening angle of the jet that produced the afterglow.
- Such analyses typically find jet opening angles in the range $\Delta\theta = 10 - 20^\circ$.
- Relating these jet opening angles to γ suggests *Lorentz factors of order 100* for many gamma-ray bursts.

Thus *afterglow lightcurve breaks* indicate directly that gamma-ray bursts are produced by *ultrarelativistic jets*, as was surmised from the preceding discussion.

21.3.4 Implications of Ultrarelativistic Beaming

The GRB beaming mechanism implies that *a fixed observer sees only a fraction of all gamma-ray bursts*.

- The ultrarelativistic nature of the jets means that the gamma-rays are highly beamed in direction.
- Afterglows are not strongly beamed after slowing, so they could be detected even for a GRB not on-axis (not aimed toward Earth).

Ultrarelativistic beaming solves a potential energy-conservation problem.

- If the energy from detected bursts were assumed to be emitted isotropically, from the energy fluxes detected on Earth total energies exceeding 10^{54} erg would be inferred for some gamma-ray bursts.
- This is *comparable to the rest mass energy of the Sun*, which would be difficult to explain by any mechanism that conserves energy.
- However, if GRBs are assumed to be emitted as collimated jets, then the total energy released would be much smaller than that inferred by an observer viewing it on-axis and assuming it to be isotropic,

This places gamma-ray bursts more in the total-energy range of well-studied events like supernova explosions.

21.3.5 Association of GRB with Galaxies

The localization provided by afterglows has permitted a number of long-period and (more recently) short-period GRB to be associated with distant galaxies.

1. Long-period (soft) bursts appear to be strongly correlated with *star-forming regions* (strong correlation with blue light in host galaxies).
2. Short-period (hard) bursts are generally fainter and sampled at smaller redshift than long-period bursts. They do not appear to be correlated with star-forming regions.
3. There is some evidence that long-period bursts are preferentially found in star-forming regions having low metallicity.

These observations provide further evidence that long-period and short-period bursts are initiated by different mechanisms.

21.3.6 Mechanisms for the Central Engine

The central engines responsible for gamma-ray bursts and the associated afterglows are not well understood, but an acceptable model for them must embody at least the following features:

1. All models require highly-relativistic jets to account for observed properties of gamma-ray bursts.
   - Lorentz γ factors of at least 200, perhaps as large as 1000, appear to be required by observations.
   - Jets focused with opening angles ~ 0.1 rad and power as large as $\sim 10^{52}$ erg
   - As will be discussed further below, long-period bursts must (at least sometimes) deliver $\sim 10^{52}$ erg to a much larger angular range (~ 1 rad) to produce an accompanying supernova, and
   - the central engine must be capable of operating for 10 seconds or longer in these long-period bursts to account for their duration.
2. The large and potentially long power timescale, particularly for long-period bursts, implies accretion onto a compact object. Thus, acceptable models must produce substantial accretion disks.

Almost the only way that we know to explain these phenomena is from a gravitational collapse and formation of an accretion disk.

One unifying idea is that gamma-ray bursts are powered by a collapse of large amounts of spinning mass to a black hole, but that there are several mechanisms to cause this. The preferred mechanisms based on current data are

1. *Particular classes of core-collapse supernovae* involving massive stars with high angular momentum for the long-period bursts.
2. *The merger of two neutron stars* or a neutron star and a black hole for the short-period bursts.

In both cases the outcome is a *Kerr black hole* having large angular momentum and strong magnetic fields, surrounded by an accretion disk of matter that has not yet fallen into the black hole.

- This scenario likely leads to highly-focused *relativistic jet outflow* on the polar axes of the Kerr black hole.
- These jets are powered by
  - rapid accretion from the disk,
  - neutrino–antineutrino annihilation,
  - strong coupling to the magnetic field.

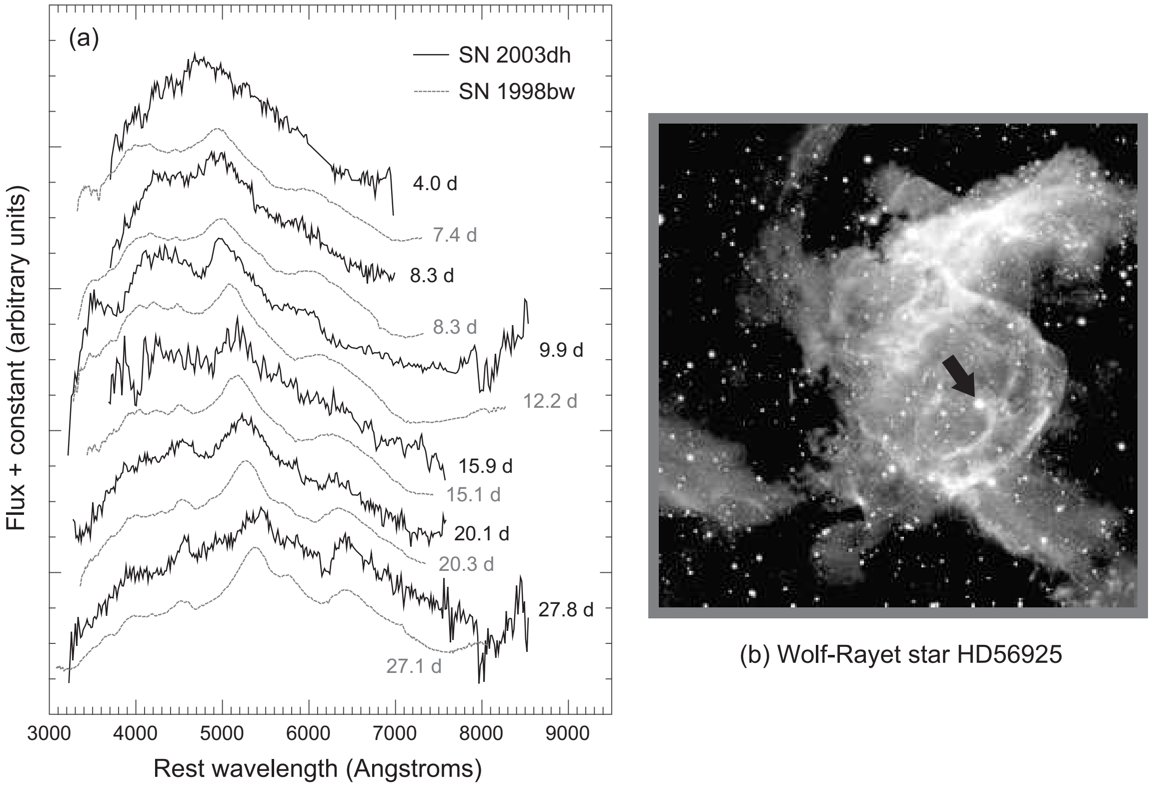
Thus, the GRB black hole engine may have many similarities with the engine powering AGN and quasars, but on a stellar rather than galactic-core scale.

Thus, we now discuss in more detail two general classes of models are now thought to account for GRB.

1. *Short-Period Bursts:* The merger of two neutron stars, or a neutron star and a black hole, with jet outflow perpendicular to the merger plane producing a burst of gamma-rays as the two objects collapse to a Kerr black hole.

2. *Long-Period Bursts:* A *hypernova,* where a spinning massive star collapses to a Kerr black hole and jet outflow from the region surrounding this collapsed object produces a burst of gamma-rays.

The unifying theme is the collapse of stellar-size amounts of spinning mass to a *Kerr black hole central engine* that powers the burst.



(b) Wolf-Rayet star HD56925

Figure 21.9: (a) Spectral bumps in the optical spectrum of SN2003dh (GRB 030329) in black, compared with a reference supernova SN1998bw in gray. The initially rather featureless spectrum of the GRB 030329 afterglow develops bumps similar to SN1998bw over time, suggesting that as the GRB afterglow fades, an underlying supernova explosion is revealed. Hence GRB 030329 also has a supernova label, SN2003dh. (b) Wolf–Rayet star (arrow) surrounded by emitted shells of gas. These massive, rapidly-spinning stars may be progenitors of Type Ib and Type Ic core collapse supernovae, and hence of long-period gamma-ray bursts.

21.3.7 Association of Long-Period GRB with Supernovae

There is a relationship between long-period gamma-ray bursts and core collapse supernovae, as suggested by Fig. 21.9.

There is an intimate relationship between *long-period GRB* and particular types of core-collapse supernovae called *Types Ib and Ic*.

- The supernova mechanism in both cases is thought to involve core collapse in a rapidly-rotating, massive (15–30 $M_\odot$) main-sequence star called a *Wolf–Rayet star.*
- These stars exhibit large mass loss and can shed their hydrogen and even helium envelopes before their cores collapse.
- They are so massive that they can collapse directly to a rotating (Kerr) black hole, instead of a neutron star.

It is thought that

- in a Type Ib supernova the H shell has been removed, and
- in a Type Ic supernova both the H and He shells have been removed

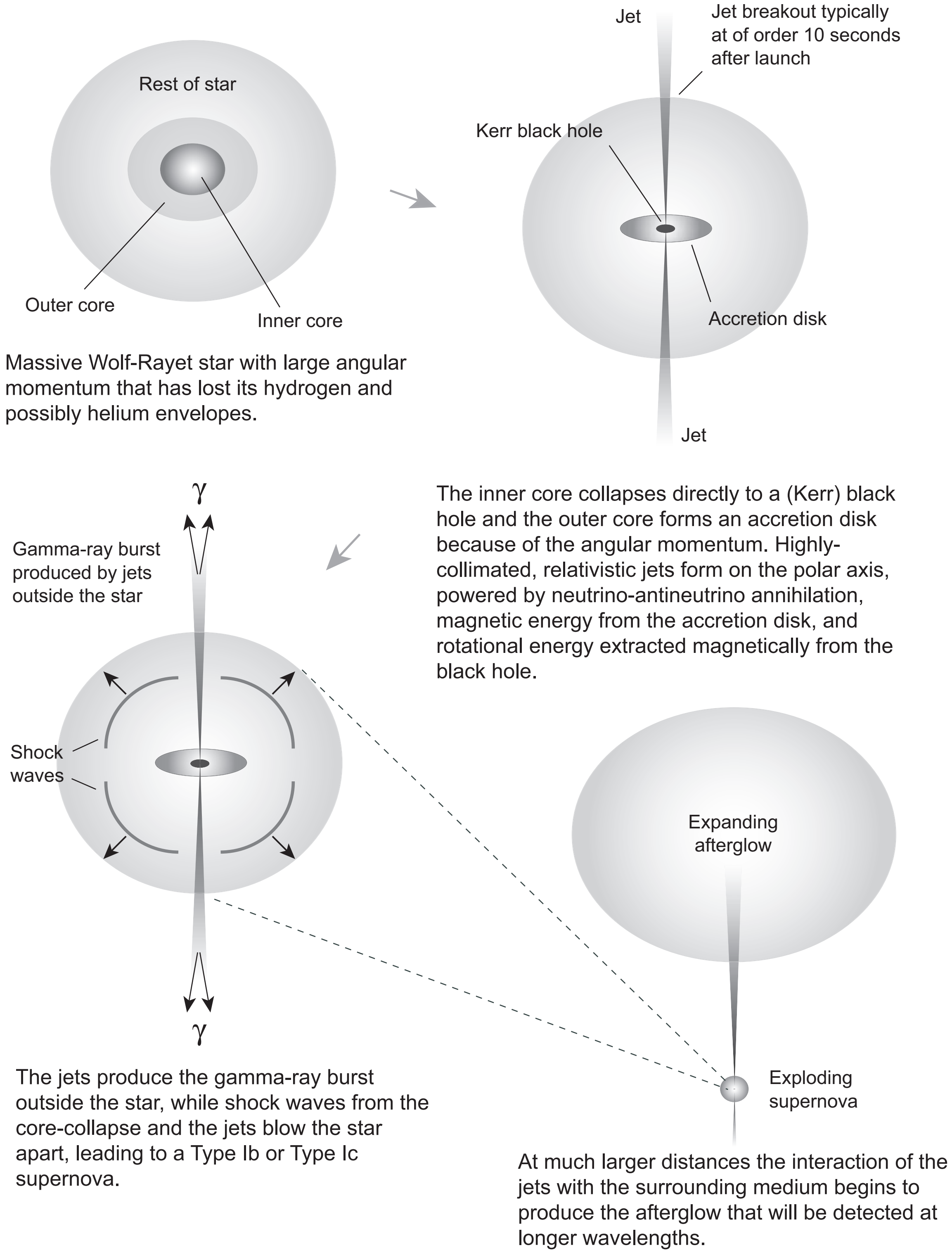
before the stellar core collapses.

Thus, long-period bursts are probably associated with core-collapse events in Wolf–Rayet stars.

- On the other hand, there is little observational evidence that short-period bursts are associated with star-forming regions, or supernovae.
- The favored mechanism for short-period bursts involves formation of an accreting Kerr black hole by merger of
  - two neutron stars, or
  - a neutron star and a black hole.

21.3.8 The Collapsar Model and Long-Period Bursts

An overview of the *collapsar model* is shown in Fig. 21.10 (next page).



Massive Wolf-Rayet star with large angular momentum that has lost its hydrogen and possibly helium envelopes.

The inner core collapses directly to a (Kerr) black hole and the outer core forms an accretion disk because of the angular momentum. Highly-collimated, relativistic jets form on the polar axis, powered by neutrino-antineutrino annihilation, magnetic energy from the accretion disk, and rotational energy extracted magnetically from the black hole.

The jets produce the gamma-ray burst outside the star, while shock waves from the core-collapse and the jets blow the star apart, leading to a Type Ib or Type Ic supernova.

At much larger distances the interaction of the jets with the surrounding medium begins to produce the afterglow that will be detected at longer wavelengths.

Figure 21.10: Collapsar model for long-period GRB and accompanying Type Ib or Ic supernova.

Simulations of relativistic jets breaking out of a Wolf–Rayet star in a collapsar model and a Wolf–Rayet star 20 seconds after core collapse are shown in Fig. 21.11 and Fig. 21.12 on the following page.

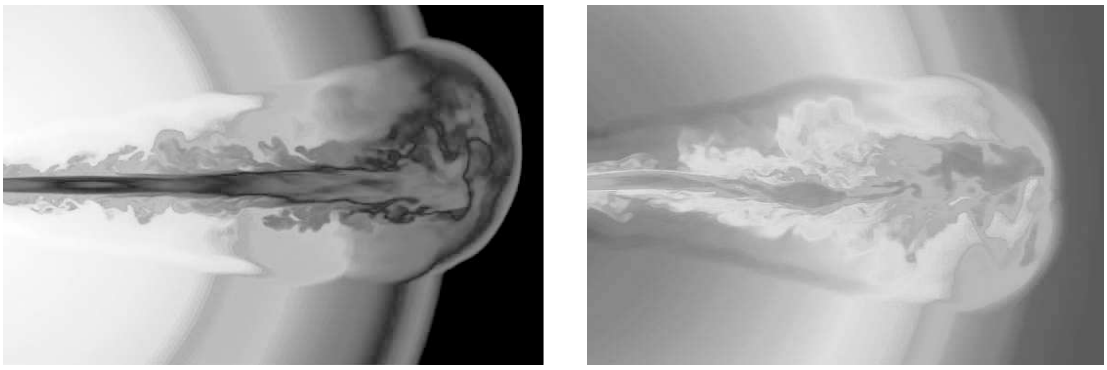


Figure 21.11: Simulations of relativistic jets breaking out of Wolf–Rayet stars. Breakout of the $\gamma \sim 200$ jet is 8 seconds after launch from the center of a 15 $M_\odot$ Wolf–Rayet star.

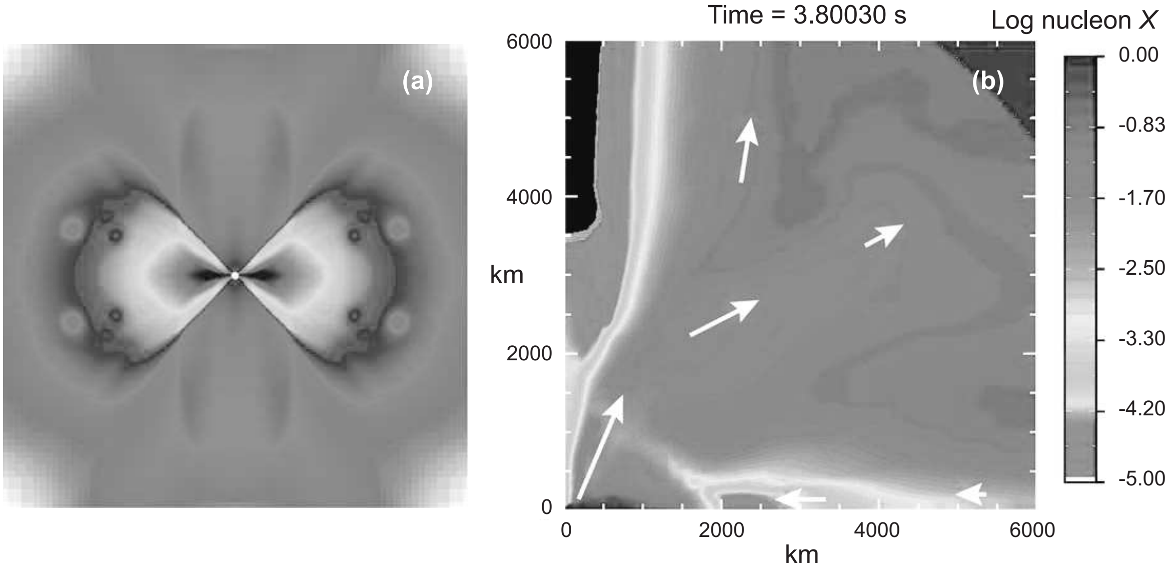
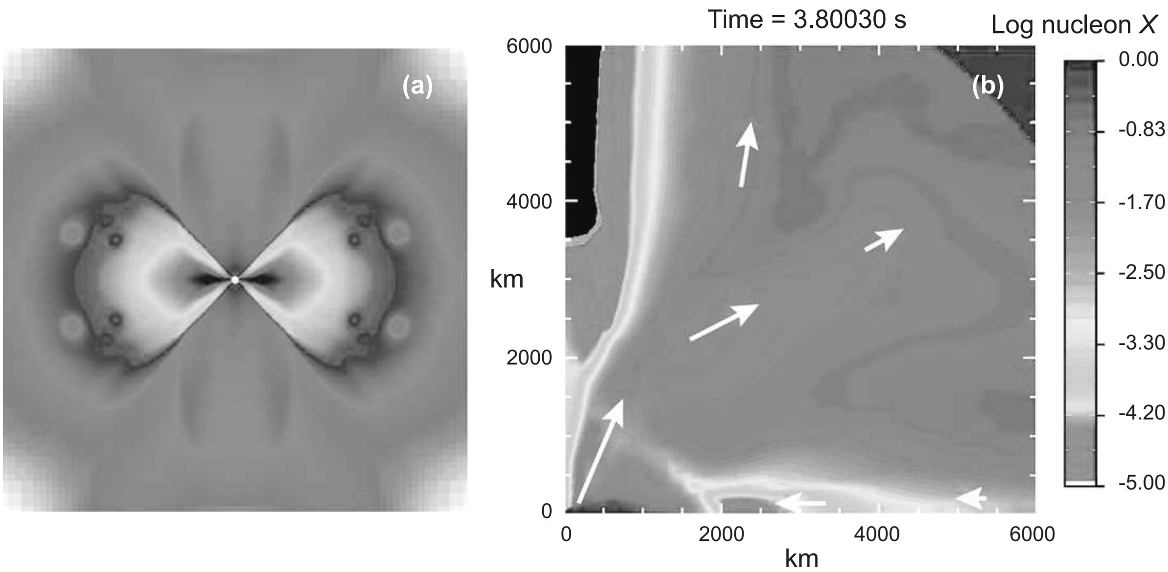


Figure 21.12: (a) A rapidly-rotating 14 $M_\odot$ Wolf–Rayet star, 20 seconds after core collapse. The polar axis is vertical, the density scale is logarithmic, and the 4.4 $M_\odot$ Kerr black hole has been accreting at $\sim 0.1 M_\odot\,\mathrm{s}^{-1}$ for 15 seconds at this point. (b) Simulation of the nucleon wind blowing off the accretion disk in a collapsar model. The gray-scale contours represent the log of the nucleon mass fraction X and arrows indicate the general flow.



In the figure above right a strong nucleon wind blowing off the collapsar accretion disk is shown. This wind

- produces the supernova and
- synthesizes the ^{56}Ni that powers the lightcurve of the supernova by radioactive decay.

The GRB and the supernova are powered in different ways in the collapsar model:

1. The GRB is powered by a relativistic jet deriving its energy from neutrino–antineutrino annihilation or rotating magnetic fields.
2. The accompanying supernova is powered by the disk wind illustrated in this figure.

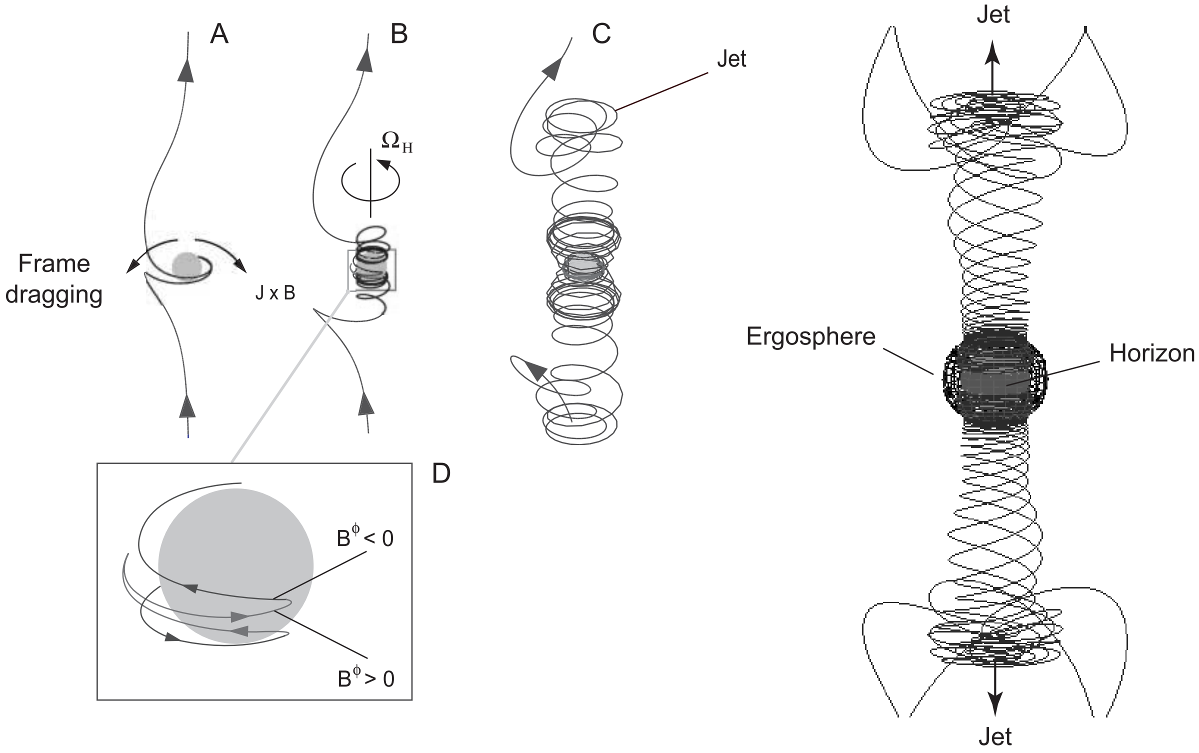


Figure 21.13: Relativistic jets produced by frame dragging of magnetic fields in the spacetime around a Kerr black hole.

Fig. 21.13 illustrates one model by which a rotating black hole could couple to a surrounding magnetic field to produce jets.

- Frame-dragging effects associated with the black hole
- wind the magnetic flux lines around the black hole and spiral them off the poles of the black hole rotation axis,
- producing bipolar ultrarelativistic jets.

The jets observed for many AGN and quasars also may be powered by a similar magnetic coupling to a Kerr black hole.

21.3.9 Merging Neutron Stars and Short-Period Bursts

The core collapse of a Wolf–Rayet star represents a plausible mechanism for long-period gamma-ray bursts that associates them naturally with star-forming regions.

- On the other hand, there is little observational evidence that short-period bursts are associated either with star-forming regions or supernovae.
- This suggests that the mechanism responsible for them must be something other than the core collapse of Wolf–Rayet stars.

The favored mechanism for short-period bursts also involves the formation of an accreting Kerr black hole, but one produced by

- the merger of two neutron stars, or
- merger of a neutron star and a black hole,

rather than by the core collapse of a massive star.

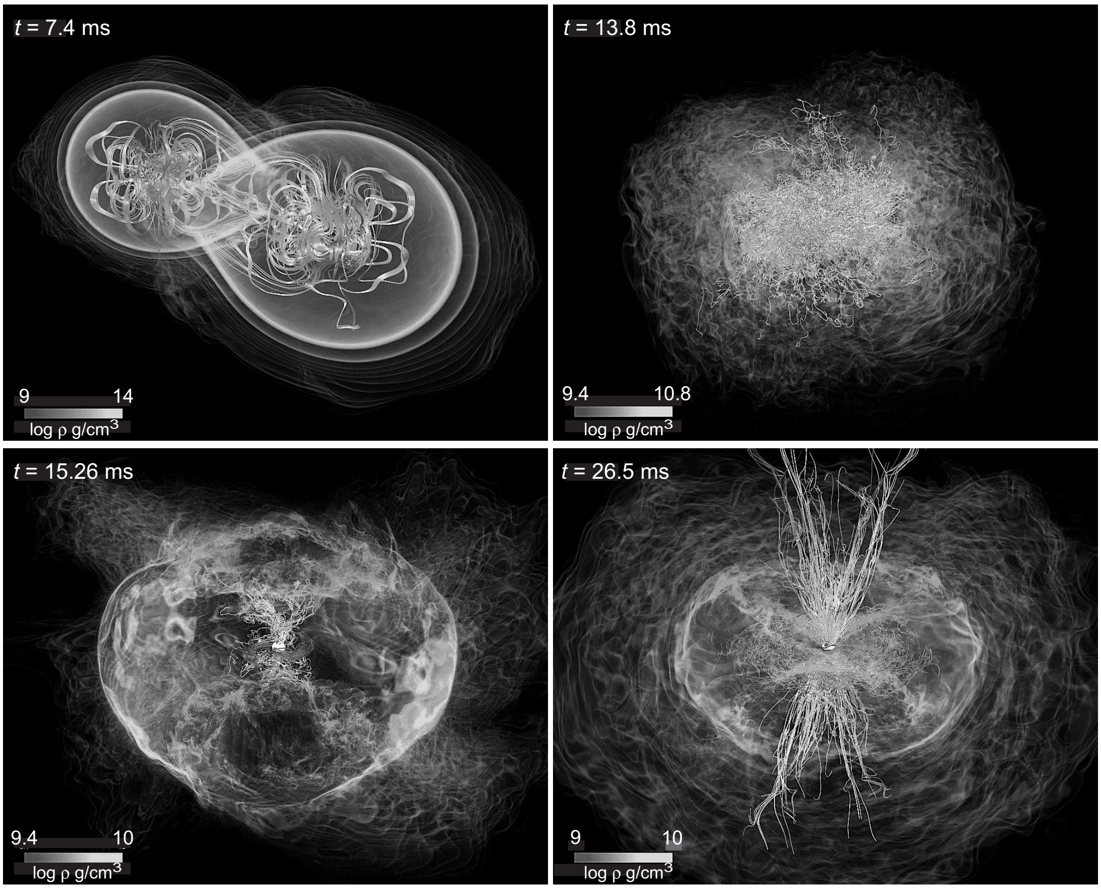


Figure 21.14: Neutron star merger simulation with strong magnetic fields.

Simulation of neutron-star merger to form a Kerr black hole with strong magnetic fields is shown in Fig. 21.14. Panels show evolution of mass density, with magnetic field lines superposed.

- The first panel shows the state shortly after initial contact.
- The second displays a merged neutron star configuration.
- In the bottom panels a Kerr black hole has formed with a disk around it, and the magnetic field is wound up by the disk to a strength $\sim 10^{15}$ gauss.

Chapter 22

Gravitational Waves and Stellar Evolution

Detection of GW150914 from merger of two black holes may be as important for stellar physics as for gravitational physics.

- The confirmation that gravitational waves exist and can be detected was a remarkable achievement for gravitational physics and for general relativity.
- But it is also arguably the most direct evidence yet for black holes, and begins to place strong new constraints on theories of massive-star evolution.
- Of even broader significance for stellar evolution was the detection in 2017 of gravitational waves from a neutron star merger in coincidence with a gamma-ray burst, accompanied by light observed at multiple wavelengths.

This chapter introduces gravitational wave astronomy and its implications for stellar evolution.

22.1 Gravitational Waves

Gravitational waves and the requisite general relativity background are covered more thoroughly in

Modern General Relativity:
Black Holes, Gravitational Waves, and Cosmology
Mike Guidry (Cambridge University Press, 2019)

This chapter will draw heavily on the discussion in that book,

- introducing only the bare minimum of mathematics and instead
- concentrating on the potential implications of gravitational wave observation for understanding stellar evolution.

It will be useful for later discussion to summarize some basic principles without getting too deeply into the mathematical weeds.

- The essential idea is that the Einstein equations introduced in Ch. 17,

  $$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R = \frac{8\pi G}{c^4}T_{\mu\nu}.$$

  admit solutions that are

  - *wavelike* and
  - *propagate at the speed of light* (or more precisely in this context, the *speed of gravity*).

- These gravitational wave solutions have many similarities with electromagnetic wave solutions of the Maxwell equations, but with some essential differences.
- The most fundamental concerns *"what is waving?"*.
  - Electromagnetic waves are propagating ripples in the electric and magnetic fields, which are defined *in spacetime*.
  - gravitational waves are ripples propagating in the metric of spacetime, so *spacetime itself,* not some field defined in spacetime, is "waving".

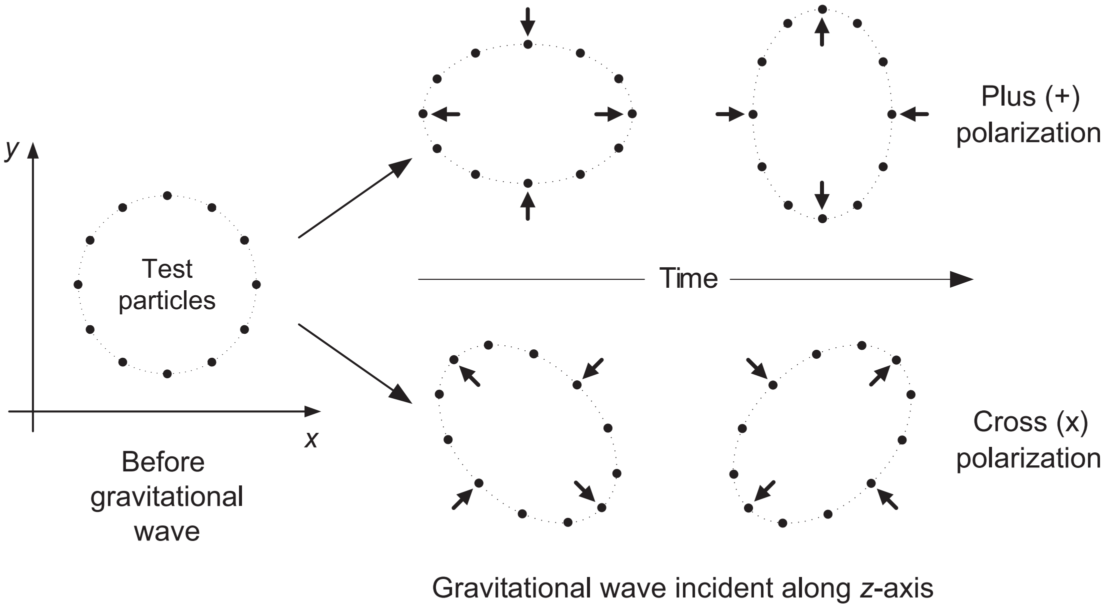

y
x
Test particles
Before gravitational wave
Time
Plus (+) polarization
Cross (x) polarization
Gravitational wave incident along z-axis

Figure 22.1: Effect of a gravitational wave incident along the z axis on test masses in the x–y plane. The top pattern is called plus $(+)$ polarization (test masses oscillate in a $+$ pattern) and the bottom pattern is called cross $(\times)$ polarization (the test masses oscillate in a $\times$ pattern).

As for electromagnetic waves,

- gravitational waves are *transverse* and
- they have *two states of polarization*.
- The two polarization states are commonly denoted *plus* $(+)$ and *cross* $(\times)$.

Gravity *acts on mass* so gravitational wave polarization may be illustrated by considering the effect of a polarized gravitational wave on a circular array of test masses, as shown in Fig. 22.1.

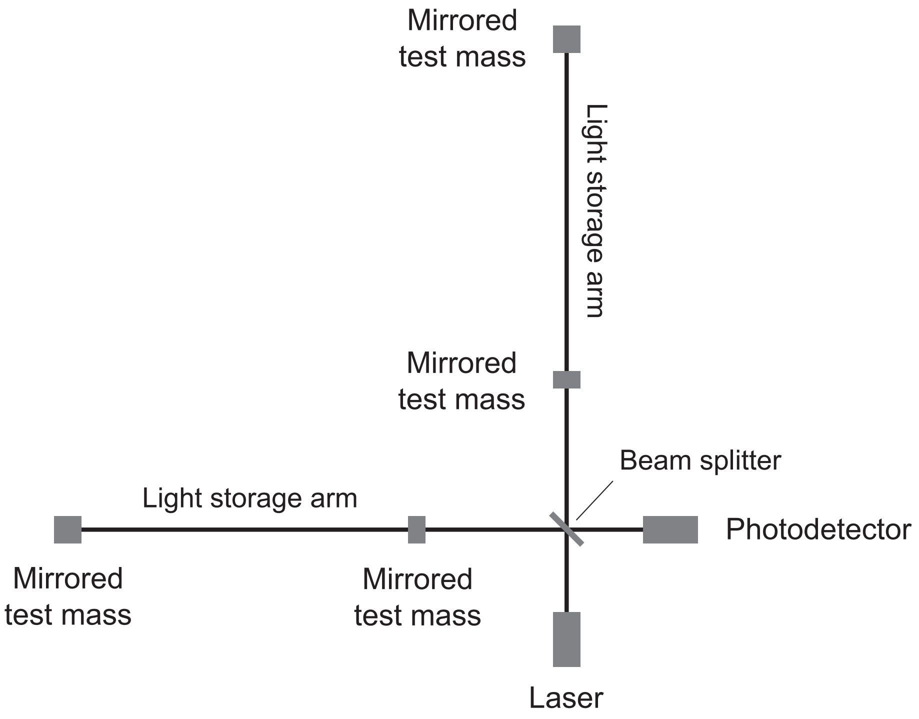


Figure 22.2: Laser interferometer gravitational wave detector. In the storage arms of actual detectors light typically is multiply reflected, greatly increasing the effective length of the arms.

These wave patterns in spacetime may be detected using Michelson laser interferometers with kilometer or longer arms, as illustrated in Fig. 22.2.

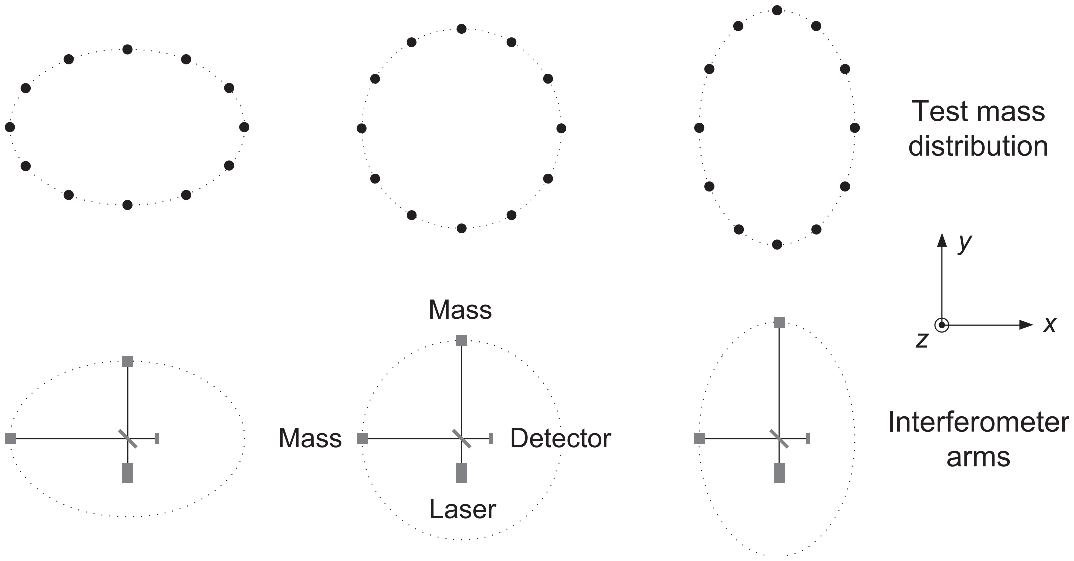
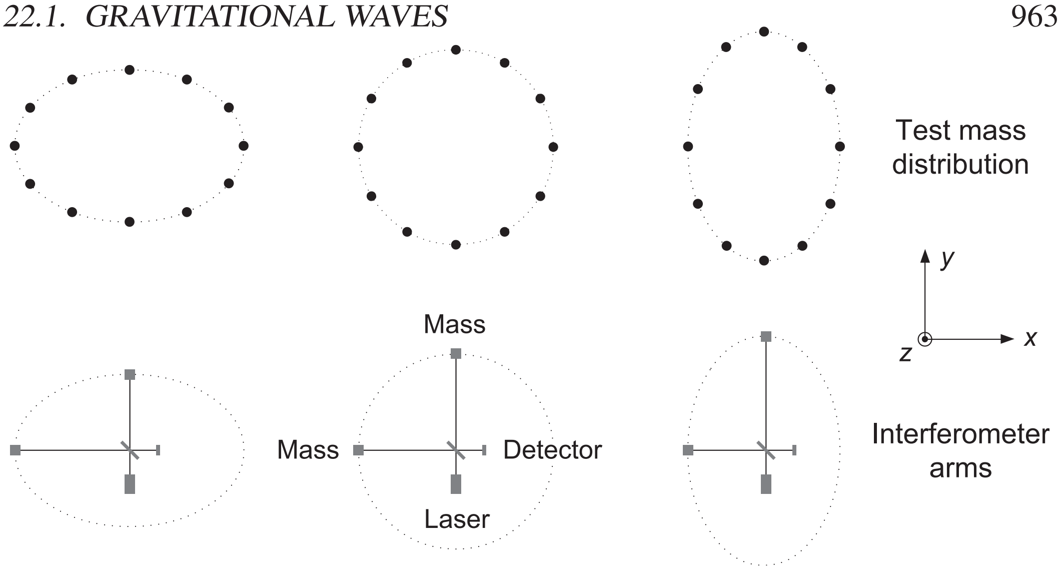


Figure 22.3: Analogy between interaction of a gravitational wave with a test mass distribution and with an interferometer.

Because the gravitational wave causes periodic fluctuations in the spacetime metric, the time for light to travel down an arm and back is modified differently for the two arms if a gravitational wave passes through the detector, as illustrated in Fig. 22.3.



By comparing the two beams,

- an interferometer can detect tiny differential changes in the light travel distances for the two arms, potentially indicating the passage of a gravitational wave.
- The fractional change in distance traveled by the light is measured by a *dimensionless strain* h, with

$$\frac{\delta L(t)}{L_0} \simeq \tfrac{1}{2} h(t,0),$$

which oscillates with the frequency of the wave.

Exquisite precision is required because

- Gravitational waves from astronomical sources require strains $h \sim 10^{-21}$ to be measured.
- $\delta L \sim hL_0$ for a strain of this size is *orders of magnitude smaller than the width of nuclei* in the atoms from which the interferometer is built!

22.2 Sample Gravitational Waveforms

We begin with an overview of some computer simulations indicating the

- varied waveforms and
- potential astrophysical information

that gravitational waves may carry.

Four kinds of events involving objects from late stellar evolution are expected to produce detectable gravitational waves:

1. Merger of two black holes,
2. Merger of a black hole and neutron star,
3. Merger of two neutron stars, and
4. A core collapse supernova explosion.

Simulations indicate that

- the corresponding waveforms carry signatures of the event that produced the gravitational wave, and that
- these may encode detailed information about the objects involved.

Some computed gravitational waveforms for events of the type described above are displayed in Figs. 22.4 (next slide).

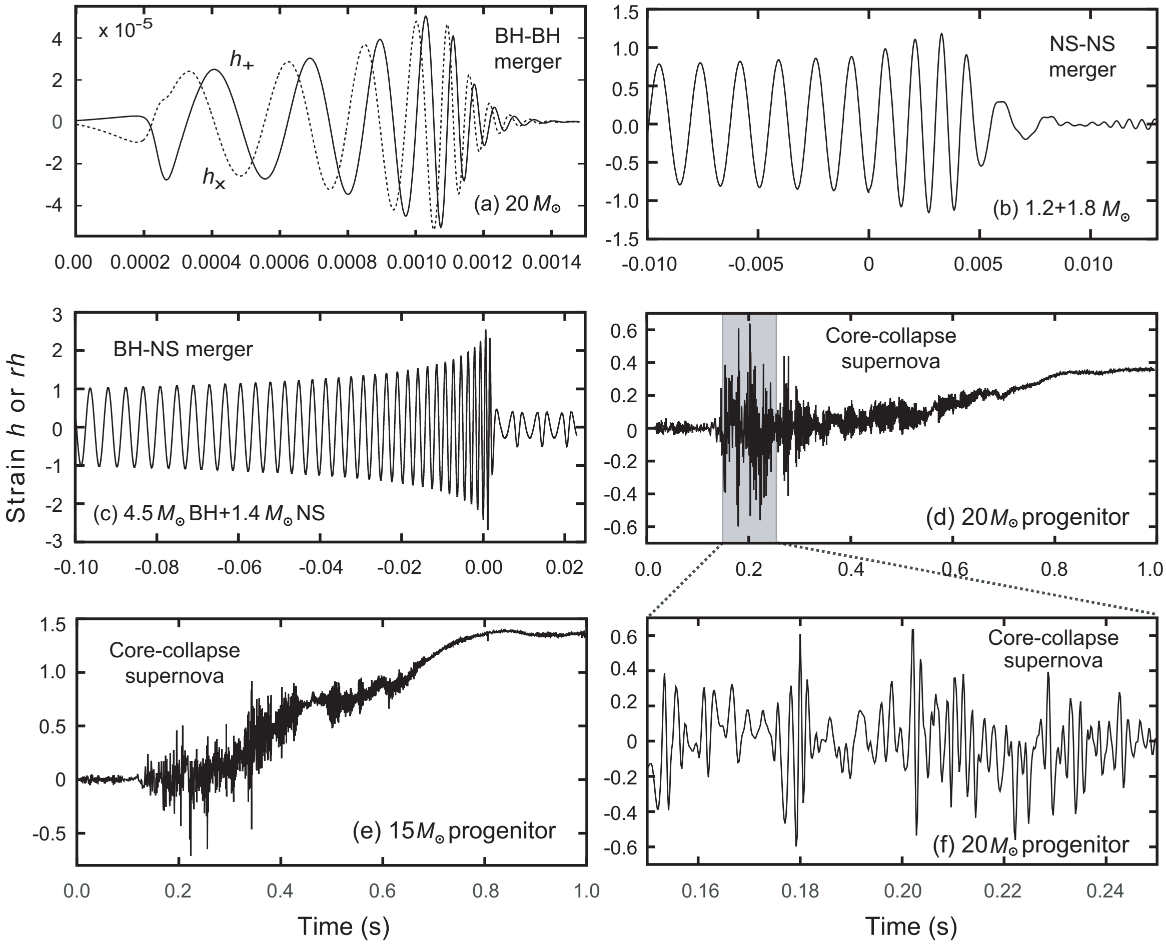
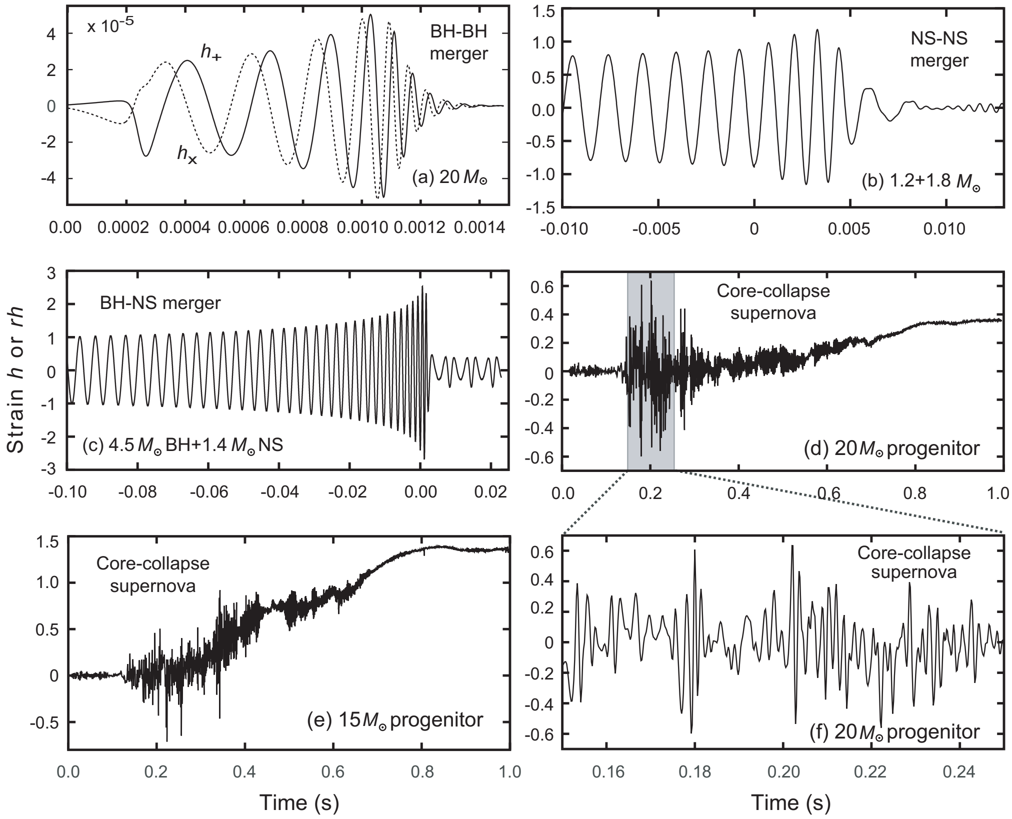


Figure 22.4: Some computed gravitational waveforms that might be observable in Earth-based detectors. (a) Merger of two 20 $M_\odot$ black holes (BH–BH). (b) Merger of 1.2 $M_\odot + 1.8$ $M_\odot$ (all masses are baryonic) neutron stars (NS–NS) at distance of 15 Mpc. (c) $4.5 M_\odot$ black hole and $1.4 M_\odot$ neutron star merger (BH–NS) at 15 Mpc. (d)–(f) Supernova at 15 kpc for two progenitor masses; time measured from bounce. Panel (f) displays the initial burst of panel (d) at higher resolution. In panel (a) rh is shown, where r is the distance to the source in cm. In panels (b)–(f) strain is given in dimensionless units of 10^{-21} by assuming a distance to the source. All waves are h_+ polarization except for in (a), where both h_+ and $h_\times$ are shown.



Different events have characteristic waveforms that are often sensitive to details:

- Mergers exhibit a *chirp waveform* (amplitude and frequency rising rapidly near merger).
- Supernova explosions are characterized by a much more complex wave pattern reflecting detailed microphysics that varies with the progenitor star.
- Hence, waveform templates may be used to identify classes of observed gravitational wave events, and
- the detailed waveforms can shed new light on the physics underlying each class of events.

As we may see by comparing these examples,

- The gravitational waveform is very dependent on the nature of the objects participating in formation of the wave.
- Hence it should be sensitive to their detailed physics.

For example, consider the following example of how gravitational waves from neutron star mergers might *constrain the equation of state* for neutron-stars.

Example: The appropriate *equation of state* to employ for neutron stars is not known very precisely.

- This is primarily because it is *difficult to measure the radius and mass simultaneously* for a neutron star.
- This introduces substantial uncertainty into the theoretical understanding of neutron stars.

Gravitational waves emitted by the merger of neutron stars would be sensitive to the properties of the neutron stars.

- This can place stronger constraints than are presently available on the neutron star equation of state.
- An improved neutron-star equation of state would permit answering more definitively questions like
  - What is the upper limit for the mass of a neutron star (which has implications for the search for black hole candidates in binary star systems)?
  - What are the superfluid and superconducting properties of neutron stars?
  - What is the relationship of observed cooling to internal structure for the neutron star?
  - Can exotic states like quark matter exist in the centers of more massive neutron stars?

Later we shall discuss observation of gravitational waves from a neutron star merger.

For 100 years after they were first proposed by Einstein, gravitational waves had been a primarily hypothetical issue, with only a few indirect observations indicating their existence. This changed dramatically in late 2015.

22.3 The Gravitational Wave Event GW150914

On September 14, 2015, almost 100 years had passed with no direct detection of gravitational waves.

- The LIGO detectors in Livingston, Louisiana and Hanford, Washington were not yet officially observing after a major upgrade, but were online and taking data.
- At 09:50:45 UTC the Livingston detector observed a strong transient lasting $\sim 200\,\text{ms}$; 7 milliseconds later the Hanford detector observed a similar transient.
- These transients were identified within 3 minutes by generic low-latency scans as *a likely gravitational wave*.
- The signal had the obvious character of a *compact merger event* (the *chirp pattern* described below).
- Low-latency data pipelines scanning with *matched filtering* quickly ruled out the merger of two neutron stars or the merger of a black hole and neutron star as the source.
- Thus attention focused almost immediately on a gravitational wave from the *coalescence of two black holes*.
- Several months of thorough analysis confirmed with greater than 5σ confidence (a false alarm probability $< 2\times 10^{-7}$) that the transient GW150914 was indeed a *gravitational wave emitted from the merger of two black holes*.

Thus the detection of GW150914 gave the *first direct confirmation* of Einstein's century-old prediction that fluctuations in the curvature of spacetime could propagate as gravitational waves.

22.3.1 Observed Waveforms

The observed waveforms in the Livingston and Hanford detectors for GW150914 are shown in Fig. 22.5 (next page)

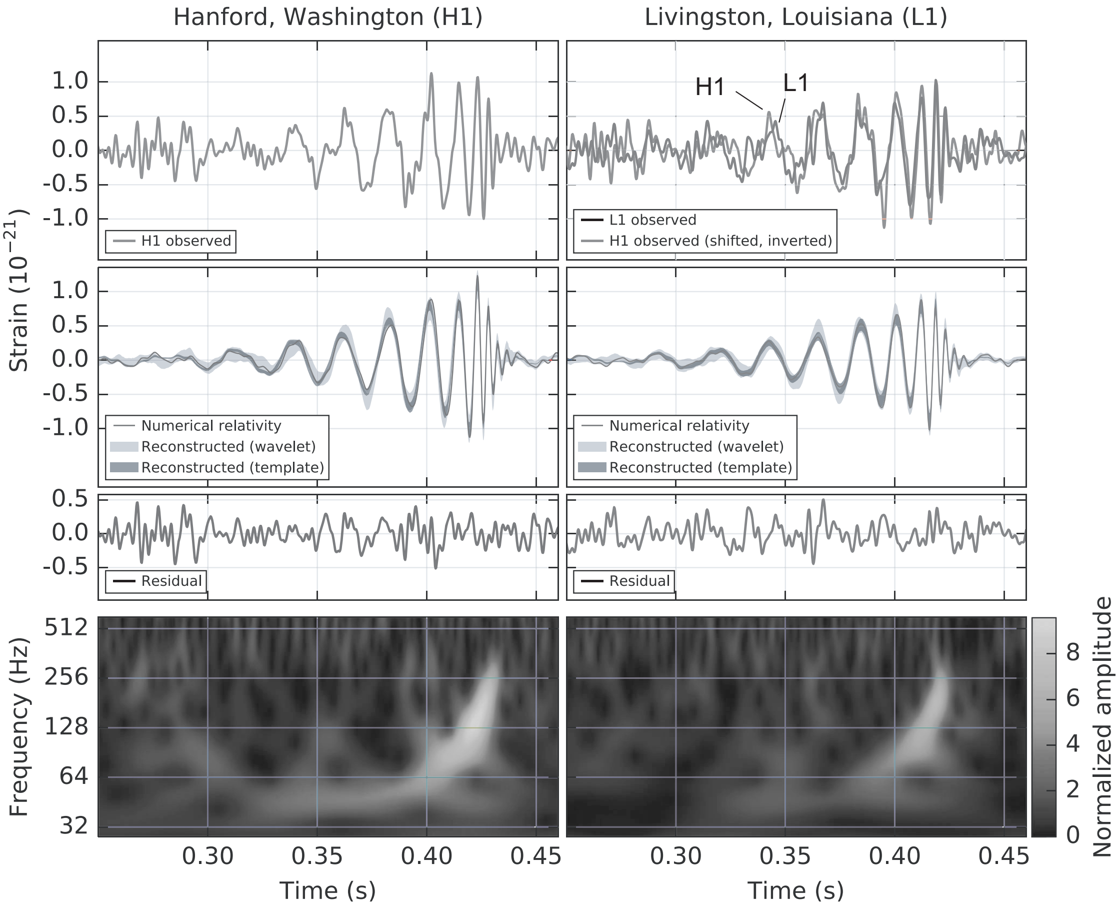
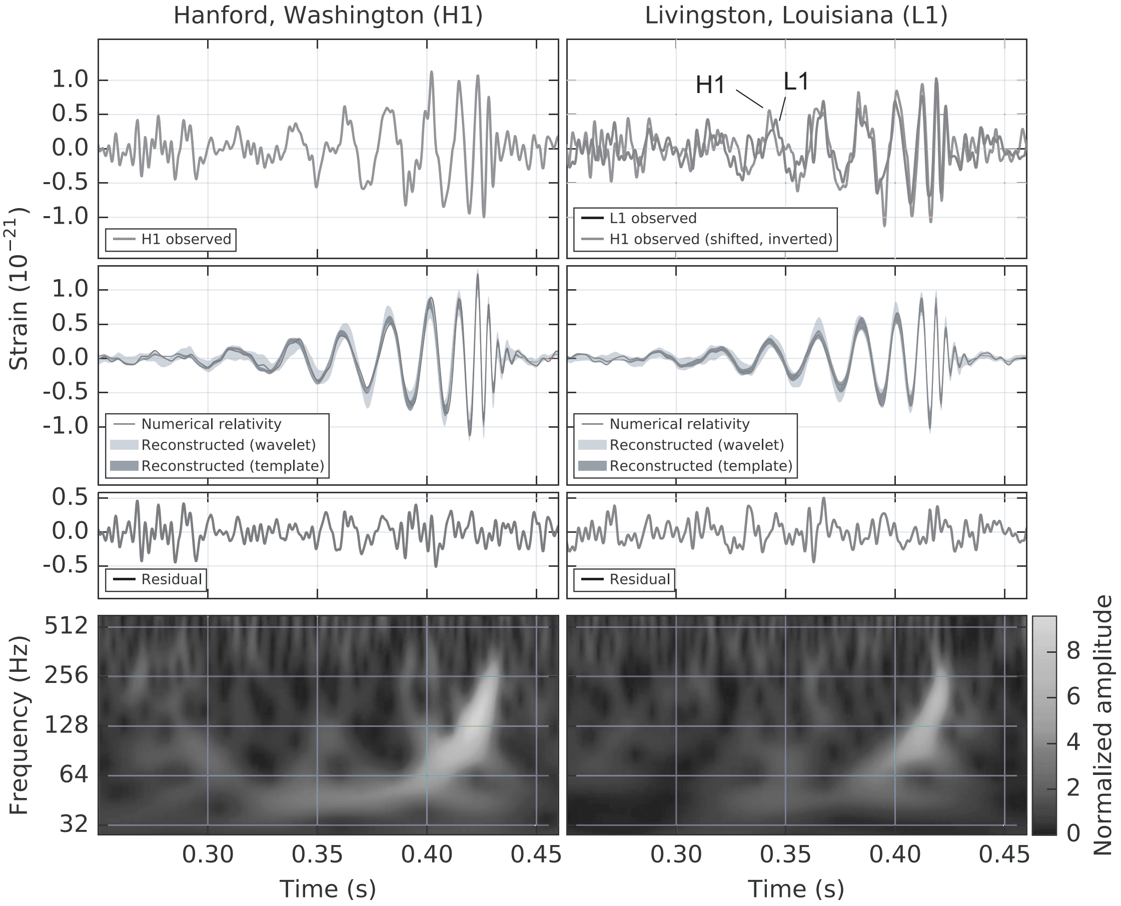
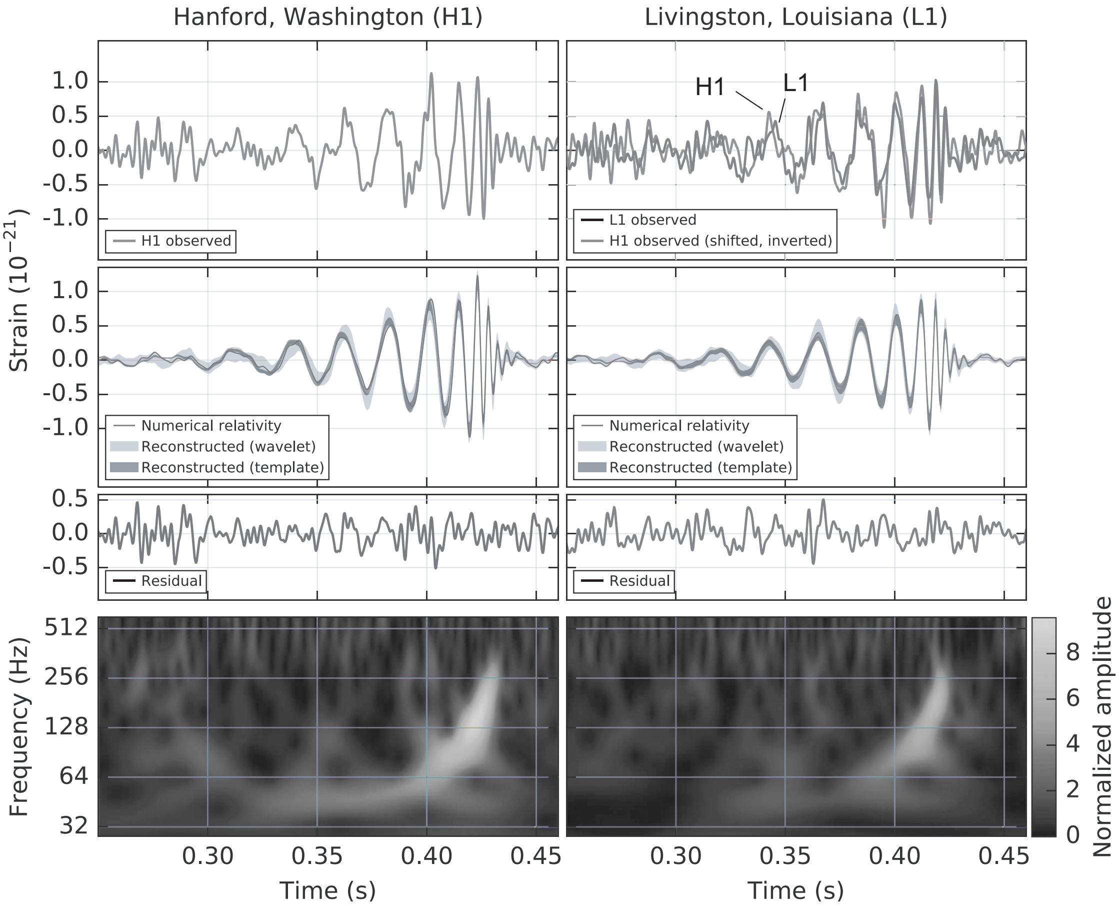


Figure 22.5: LIGO gravitational wave event GW150914. Left panels correspond to data from the Hanford detector (H1) and right panels to data from the Livingston detector (L1). Top row is measured strain in units of 10^{-21}. In the top right panel the Hanford signal has been superposed on the Livingston signal. The second row shows numerical relativity simulations of the waveform assuming a binary black hole merger event. The third row shows residuals after subtracting the numerical relativity waveform (second row) from the detector waveform (first row). The fourth row shows frequency versus time for the strain data, with grayscale contours indicating strain amplitude. The rapidly-rising pattern (chirp) is indicative of a binary merger.



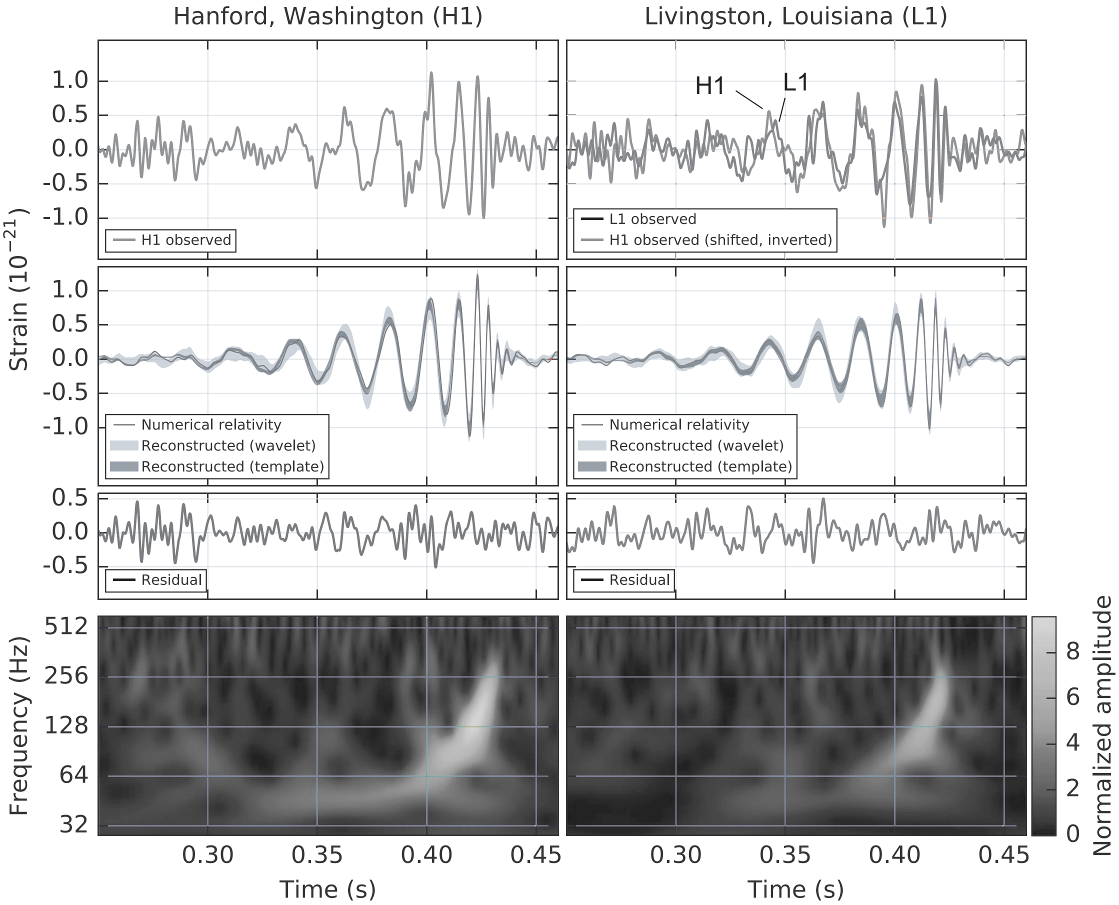
- The gravitational wave arrived first at Livingston (L1) and then $6.9^{+0.5}_{-0.4}$ ms later at Hanford (H1).
- In the top-right image the H1 wave has been superposed, shifted by 6.9 ms, and inverted to account for relative orientations of the two detectors (orientations relative to local north of L1 and H1 differ by 72°).
- This superposition and a 24:1 signal to noise ratio leaves little doubt that the same wavefront, traveling at light-speed, passed first through Livingston and then Hanford.



- In the second row of the above figure, numerical relativity simulations of the waveform for merging black holes and wavelet reconstructions with and without an astrophysical black hole merger model are shown.

- The third row displays the result of subtracting the numerical relativity waveform in the second row from the observed waveform in the first row.

- The last row shows a *time-frequency representation of the data*, with the grayscale contours representing strain.



- The frequency–time plot in the bottom row indicates that over a period of ~ 0.2 seconds the signal swept upward in frequency from about 35 Hz to 250 Hz.

- This signal, rising in frequency and strain (*“the chirp”*), is indicative of the final rapid inspiral of a merger event, with a peak strain $\sim 1.0 \times 10^{-21}$.

- It may be noted that this strain changed the separation between the test masses by only about 4×10^{-16} cm, which is *0.005 times the diameter of a proton*.

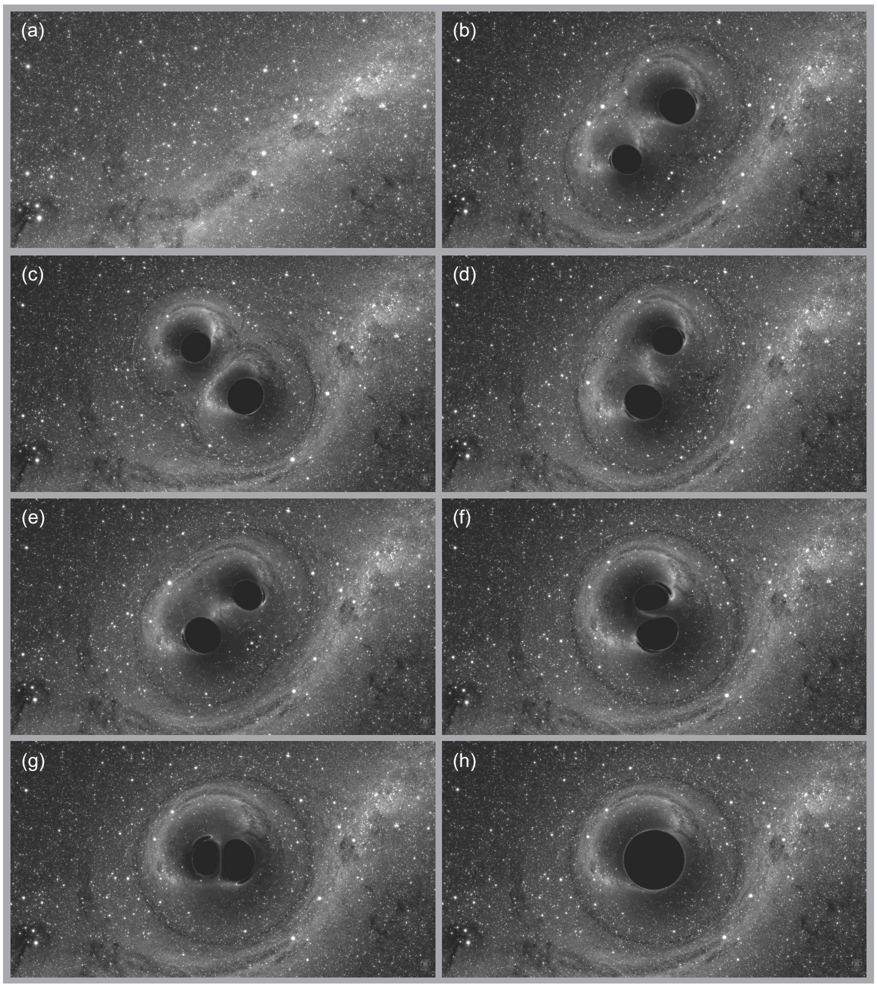


Figure 22.6: Computer simulation of the GW150914 merger. Panel (a) is the undisturbed background field of stars. (b)–(h) are succesively later frames.

In Fig. 22.6 a simulation of what the black holes might have looked like from up close during the merger is shown.

- The dark, well-defined shapes are the shadows of the black hole event horizons as they block all light from behind.
- The flattened dark features around them and distorted star fields are strong gravitational lensing effects.

Table 22.1: The black-hole merger event GW150914

Quantity	Value[†]
Primary black hole mass	$36^{+5}_{-4}\ M_\odot$
Secondary black hole mass	$29^{+4}_{-4}\ M_\odot$
Final black hole mass	$62^{+4}_{-4}\ M_\odot$
Final black hole spin	$0.67^{+0.05}_{-0.07}$
Mass radiated as gravitational waves	$3.0^{+0.5}_{-0.5}\ M_\odot$
Peak gravitational wave luminosity (erg s^{-1})	$3.6^{+0.5}_{-0.4} \times 10^{56}$
Peak gravitational wave luminosity ($M_\odot\ \mathrm{s}^{-1}$)	200^{+30}_{-20}
Source redshift z	$0.09^{+0.03}_{-0.04}$
Source luminosity distance	410^{+160}_{-180} Mpc

[†] Masses in source frame. Multiply by $(1+z)$, where z is redshift, for mass in detector frame. Spin given in units of spin for an extreme Kerr black hole of that mass.

Extensive analysis comparing simulations of the merger with data measured for the gravitational wave yields quantitative information about

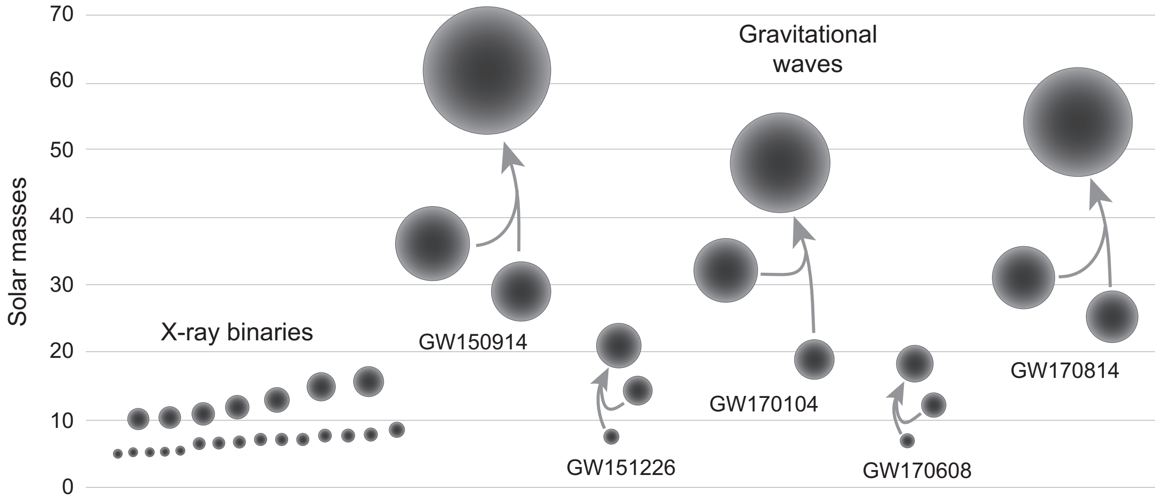
- the *two black holes that merged*, and
- the *final Kerr black hole* that resulted from the merger.

These parameters for GW150914 are displayed in Table 22.1.

- The initial masses of the merging black holes were determined to be $36\,M_\odot$ and $29\,M_\odot$, respectively.
- The mass of the final black hole was $62\,M_\odot$.
- Thus from the difference of initial and final masses, about $3\,M_\odot$ was radiated as gravitational waves.
- The 3 solar masses were converted to gravitational waves over a period of less than half a second.
- This corresponded to a peak gravitational wave luminosity of an astonishing $\sim 200\,M_\odot\,\mathrm{s}^{-1}$!
- This translates through $E = mc^2$ to a peak luminosity of well over $10^{56}\,\mathrm{erg\,s}^{-1}$.
- This peak luminosity is some
  - 23 orders of magnitude greater than the Sun's photon luminosity and
  - 5 orders of magnitude brighter than the photon luminosity of a supernova.
- The redshift and corresponding distance to the source were $z = 0.09$ and $410\,\mathrm{Mpc}$, respectively.
- The spin of the final black hole was determined to be 67% of that for an extremal Kerr black hole.

The direction to the source was determined also.

- Since the gravitational wave was observed by only two detectors, tracking the wave back to its source entailed considerable uncertainty.
- The analysis was able to localize the source to an error box of about $230\,\mathrm{deg}^2$ in the Southern Hemisphere near the Large Magellanic Cloud.
- As more gravitational wave observatories come online and a signal can be triangulated from more than two detectors, this uncertainty will decrease.
- For example, the more recent gravitational wave event GW170817 to be discussed shortly was localized to $28\,\mathrm{deg}^2$.
- However, gravitational wave detectors will always have *lower intrinsic angular resolution* than traditional astronomy instruments.
- On the other hand, gravitational wave interferometers *see essentially the entire sky at all times*, not just a narrow field as for traditional telescopes.



22.4 A New Probe of Massive-Star Evolution

Notice from Fig. 17.7, reproduced above, that

- Each of the two initial black holes for GW150914 had at least a factor of two more mass than the most massive black holes inferred from X-ray binary data.
- Thus GW150914 provided the first conclusive evidence
  - that such massive black holes can exist,
  - that they can occur in binary pairs, and
  - that these binaries can form with sufficiently compact orbits that they can merge within the age of the Universe through gravitational wave emission.

Understanding this has implications for the evolution of massive stars, in particular for those in binary systems.

22.4.1 Formation of Massive Black Hole Binaries

The formation of massive black hole binaries implied by the merger event GW150914 requires a sequence of four events to occur in the course of stellar evolution.

1. Stars must form with very large masses (probably in the vicinity of $100 M_\odot$).
2. These stars must not lose too much of their mass to stellar winds while evolving to core collapse.
3. These massive stars must collapse to black holes, so they must avoid
   - *collapsing to neutron stars* and
   - *destruction by the pair instability* discussed in the supernova chapter.
4. The black holes thus formed must end up as part of a binary star system.

Thus the interpretation of GW150914 as resulting from merger of two $\sim 30 M_\odot$ black holes poses some *challenging questions* for theories of stellar evolution.

22.5 Gravitational Waves and Stellar Evolution

How does a binary composed of $30M_\odot$ black holes even form? Presumably either

- A binary formed with two stars of large mass and *survived successive core collapses* for each star, or
- The black holes *formed independently* through core collapse of massive stars in a dense cluster and then were *captured by mutual gravity* into a binary orbit.

Neither scenario is easy to realize without assumptions that are not well tested. Therefore, we may expect that future detection of gravitational wave events from

- merger of two black holes,
- merger of a black hole and neutron star,
- merger of two neutron stars, and
- core collapse supernovae

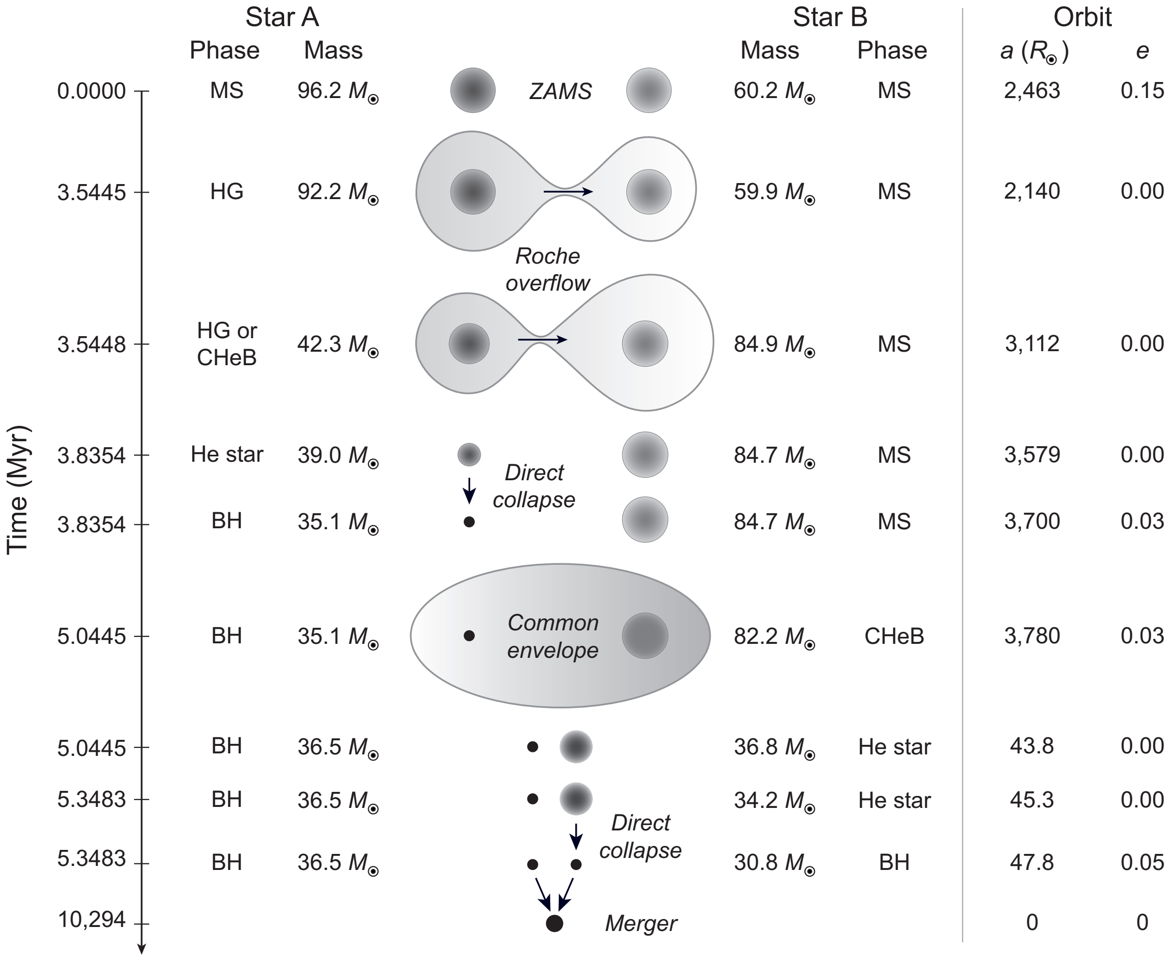
will shed considerable light on—and pose substantial challenges to—our general understanding of stellar evolution, particularly for the neutron star and black hole endpoints for massive-stars evolution.

Comprehensive simulations indicate that the binary black holes responsible for the gravitational waves observed thus far by LIGO

- could have formed in isolated binary star evolution,
- provided that they formed in regions having low concentration of elements heavier than helium (regions of *low metallicity*).

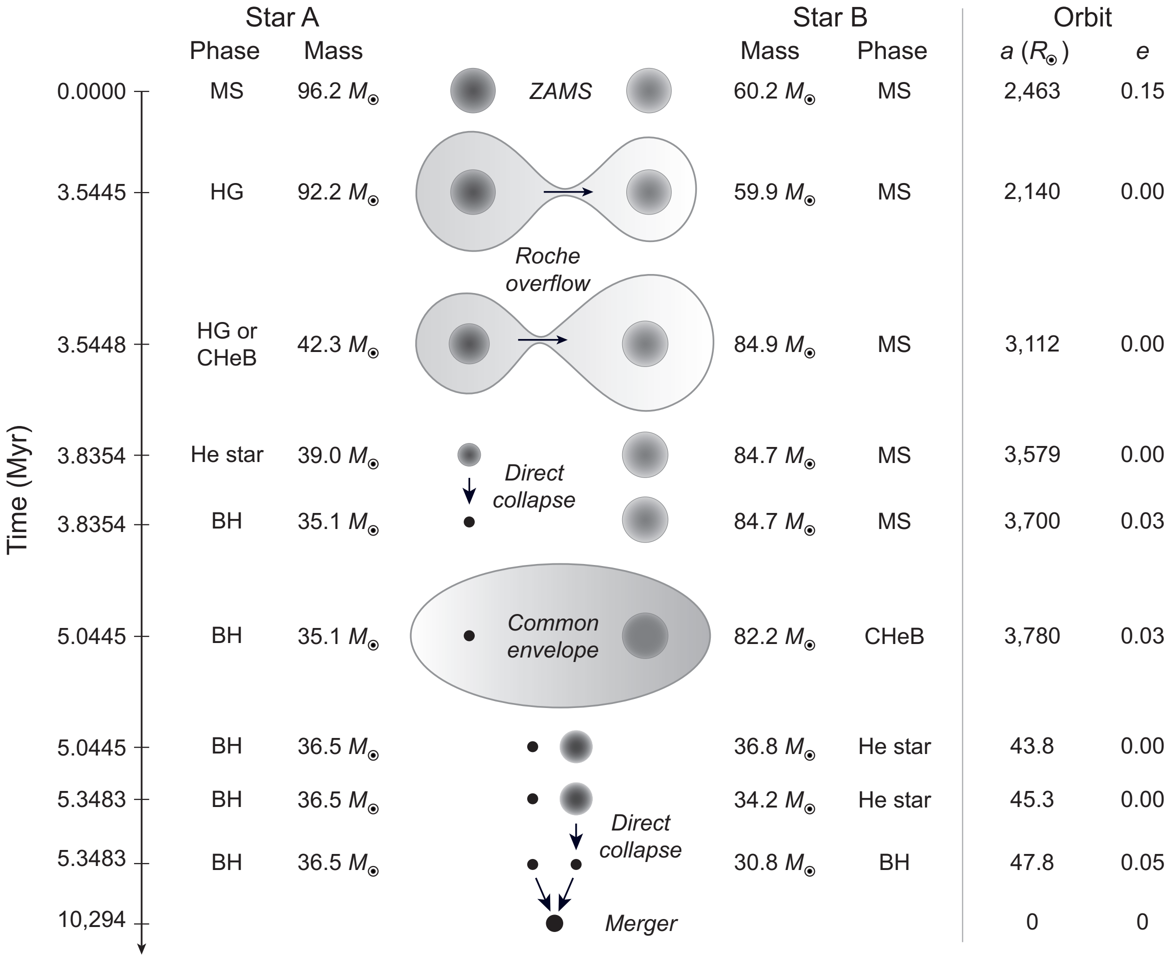
22.5.1 A Possible Evolutionary Scenario for GW150914

Figure 22.7 (following page) illustrates a possible scenario for the production of GW150914.



	Star A		Star B		Orbit	
Time (Myr)	Phase	Mass	Mass	Phase	a ($R_\odot$)	e
0.0000	MS	96.2 $M_\odot$	60.2 $M_\odot$	MS	2,463	0.15
3.5445	HG	92.2 $M_\odot$	59.9 $M_\odot$	MS	2,140	0.00
3.5448	HG or CHeB	42.3 $M_\odot$	84.9 $M_\odot$	MS	3,112	0.00
3.8354	He star	39.0 $M_\odot$	84.7 $M_\odot$	MS	3,579	0.00
3.8354	BH	35.1 $M_\odot$	84.7 $M_\odot$	MS	3,700	0.03
5.0445	BH	35.1 $M_\odot$	82.2 $M_\odot$	CHeB	3,780	0.03
5.0445	BH	36.5 $M_\odot$	36.8 $M_\odot$	He star	43.8	0.00
5.3483	BH	36.5 $M_\odot$	34.2 $M_\odot$	He star	45.3	0.00
5.3483	BH	36.5 $M_\odot$	30.8 $M_\odot$	BH	47.8	0.05
10,294					0	0

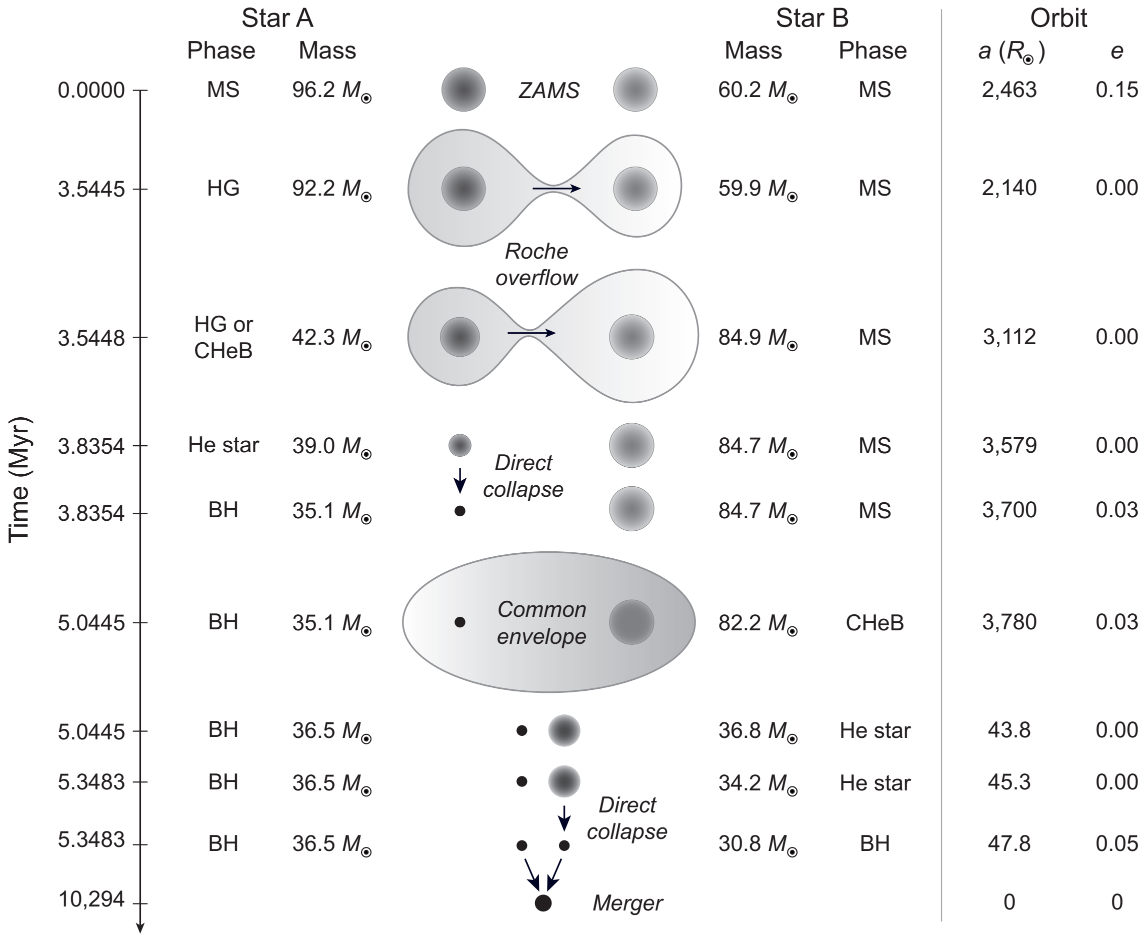
Figure 22.7: A scenario for evolution of the massive black hole binary leading to GW150914. ZAMS means zero age main sequence (the time when the star first enters the main sequence), MS means main sequence, HG means a star evolving through the Hertzsprung gap (the evolutionary region between the main sequence and the red giant branch), CHeB means core helium burning, a He star is a star exhibiting strong He and weak H lines (indicating loss of much of its outer envelope), and BH indicates a black hole. Time is measured from formation of the binary and the scale is *highly nonlinear*. The separation of the pair is a and the eccentricity of the orbit is e.



	Star A		Star B		Orbit	
Time (Myr)	Phase	Mass	Mass	Phase	a ($R_\odot$)	e
0.0000	MS	96.2 $M_\odot$	60.2 $M_\odot$	MS	2,463	0.15
3.5445	HG	92.2 $M_\odot$	59.9 $M_\odot$	MS	2,140	0.00
3.5448	HG or CHeB	42.3 $M_\odot$	84.9 $M_\odot$	MS	3,112	0.00
3.8354	He star	39.0 $M_\odot$	84.7 $M_\odot$	MS	3,579	0.00
3.8354	BH	35.1 $M_\odot$	84.7 $M_\odot$	MS	3,700	0.03
5.0445	BH	35.1 $M_\odot$	82.2 $M_\odot$	CHeB	3,780	0.03
5.0445	BH	36.5 $M_\odot$	36.8 $M_\odot$	He star	43.8	0.00
5.3483	BH	36.5 $M_\odot$	34.2 $M_\odot$	He star	45.3	0.00
5.3483	BH	36.5 $M_\odot$	30.8 $M_\odot$	BH	47.8	0.05
10,294					0	0

A massive binary formed about 2 billion years after the big bang (redshift $z \sim 3.2$), with

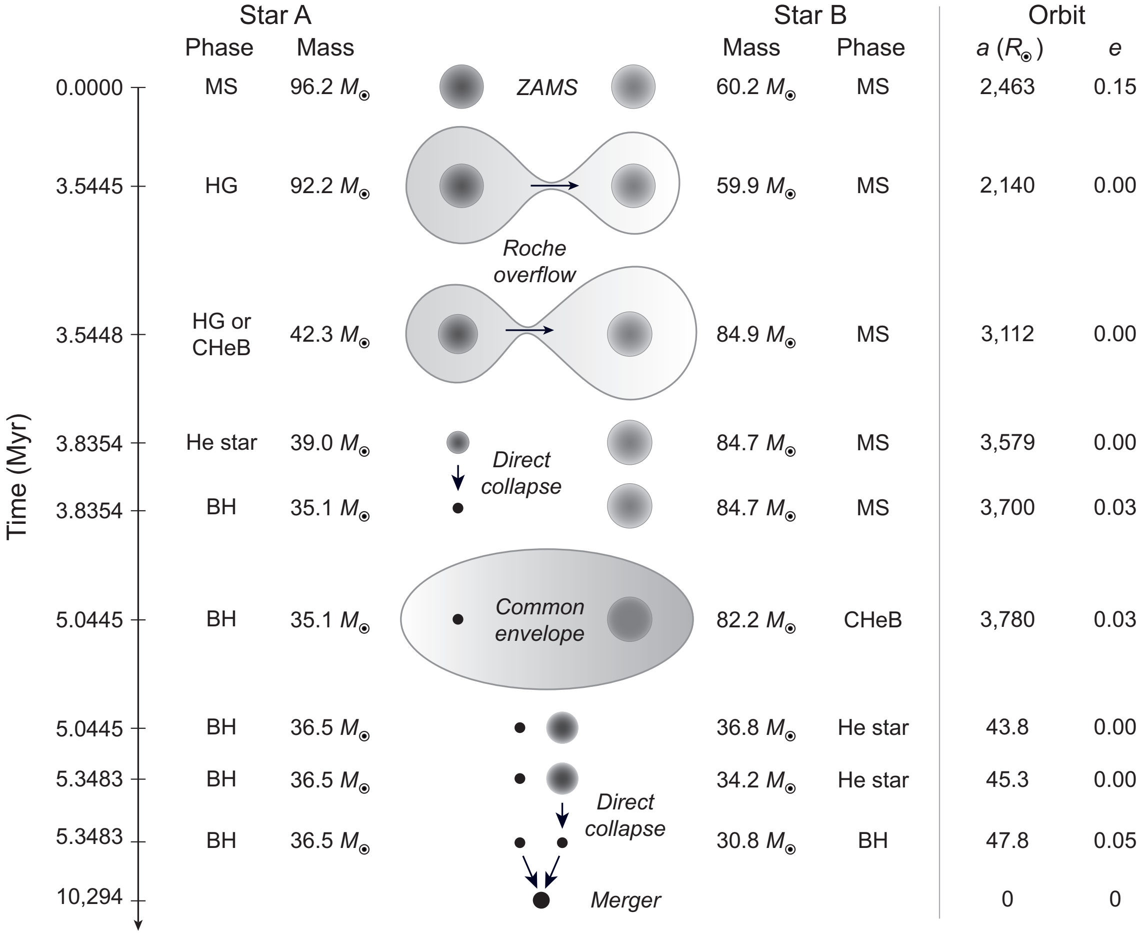
- initial main sequence masses of $96.2\,M_\odot$ (star A) and $60.2\,M_\odot$ (star B),
- a metal fraction Z that was 0.03 times that of the Sun,
- an average separation $a \sim 2500\,R_\odot$, and
- an orbital eccentricity $e = 0.15$.



Time (Myr)	Star A Phase	Star A Mass	Star B Mass	Star B Phase	Orbit a ($R_\odot$)	Orbit e
0.0000	MS	96.2 $M_\odot$	60.2 $M_\odot$	MS	2,463	0.15
3.5445	HG	92.2 $M_\odot$	59.9 $M_\odot$	MS	2,140	0.00
3.5448	HG or CHeB	42.3 $M_\odot$	84.9 $M_\odot$	MS	3,112	0.00
3.8354	He star	39.0 $M_\odot$	84.7 $M_\odot$	MS	3,579	0.00
3.8354	BH	35.1 $M_\odot$	84.7 $M_\odot$	MS	3,700	0.03
5.0445	BH	35.1 $M_\odot$	82.2 $M_\odot$	CHeB	3,780	0.03
5.0445	BH	36.5 $M_\odot$	36.8 $M_\odot$	He star	43.8	0.00
5.3483	BH	36.5 $M_\odot$	34.2 $M_\odot$	He star	45.3	0.00
5.3483	BH	36.5 $M_\odot$	30.8 $M_\odot$	BH	47.8	0.05
10,294					0	0

Star A evolved quickly, expanded, and *transferred more than half of its mass to star B* by Roche lobe overflow, as star A evolved through the Hertzsprung gap to core helium burning.

- Star A then *collapsed directly to a* 35.1$M_\odot$ *black hole*, with no ejection of a supernova remnant, but with *10% of the mass carried off by neutrinos* during the collapse.
- By the time the first black hole had formed, star B had grown by accretion to 84.7$M_\odot$ and it evolved quickly off the main sequence to core helium burning.

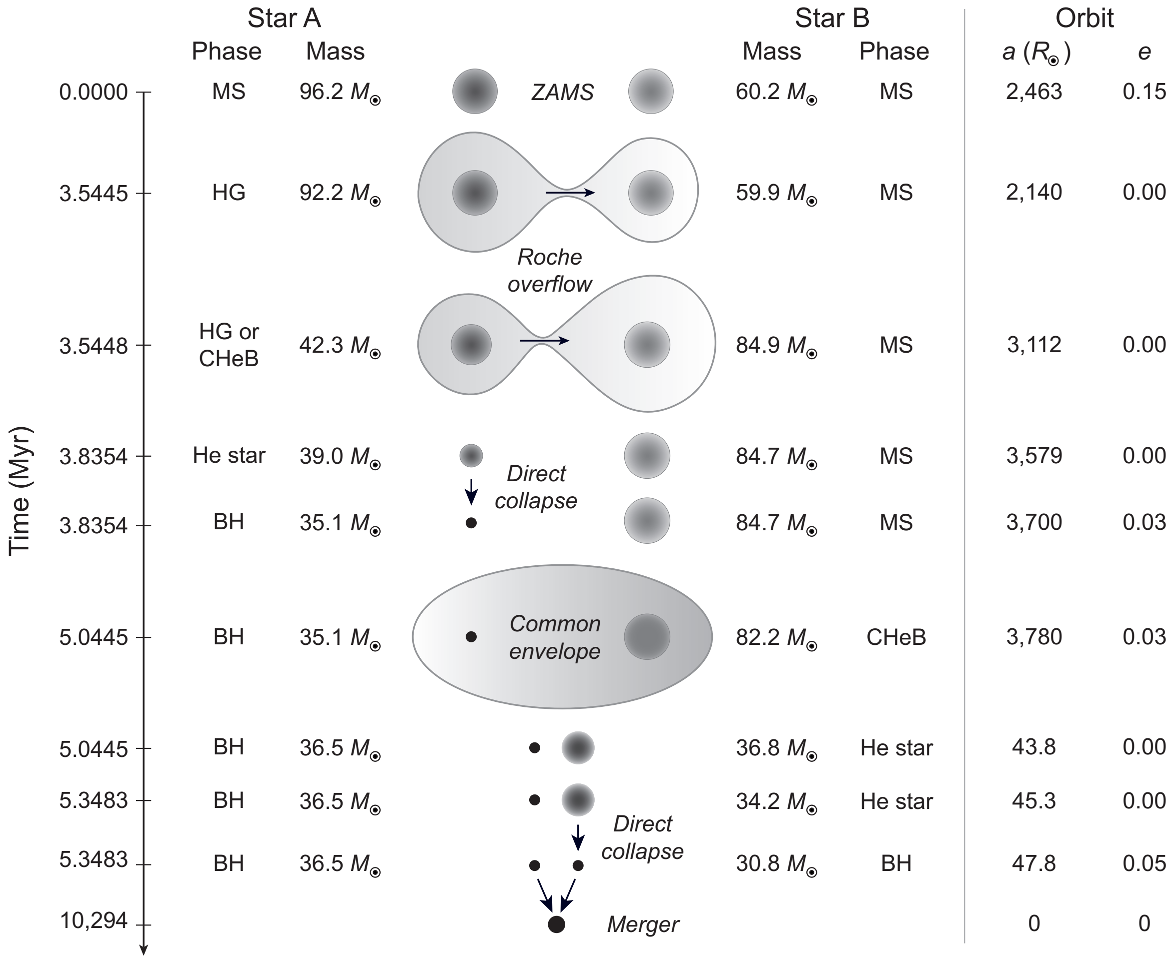


Time (Myr)	Star A Phase	Star A Mass	Star B Mass	Star B Phase	Orbit a ($R_\odot$)	Orbit e
0.0000	MS	96.2 $M_\odot$	60.2 $M_\odot$	MS	2,463	0.15
3.5445	HG	92.2 $M_\odot$	59.9 $M_\odot$	MS	2,140	0.00
3.5448	HG or CHeB	42.3 $M_\odot$	84.9 $M_\odot$	MS	3,112	0.00
3.8354	He star	39.0 $M_\odot$	84.7 $M_\odot$	MS	3,579	0.00
3.8354	BH	35.1 $M_\odot$	84.7 $M_\odot$	MS	3,700	0.03
5.0445	BH	35.1 $M_\odot$	82.2 $M_\odot$	CHeB	3,780	0.03
5.0445	BH	36.5 $M_\odot$	36.8 $M_\odot$	He star	43.8	0.00
5.3483	BH	36.5 $M_\odot$	34.2 $M_\odot$	He star	45.3	0.00
5.3483	BH	36.5 $M_\odot$	30.8 $M_\odot$	BH	47.8	0.05
10,294					0	0

The expansion of star B initiated a *common envelope (CE) phase* with the black hole that formed from star A.

- During the CE phase the average separation of the binary components was reduced from $a \sim 3800\,R_\odot$ to $a \sim 45\,R_\odot$.
- At the end of the CE phase the mass of the black hole formed from star A was $36.5\,M_\odot$ and star B was now a helium star of mass $36.8\,M_\odot$.

Star B then *collapsed directly to a black hole*.



	Star A		Star B		Orbit	
	Phase	Mass	Mass	Phase	a ($R_\odot$)	e
0.0000	MS	96.2 $M_\odot$	60.2 $M_\odot$	MS	2,463	0.15
3.5445	HG	92.2 $M_\odot$	59.9 $M_\odot$	MS	2,140	0.00
3.5448	HG or CHeB	42.3 $M_\odot$	84.9 $M_\odot$	MS	3,112	0.00
3.8354	He star	39.0 $M_\odot$	84.7 $M_\odot$	MS	3,579	0.00
3.8354	BH	35.1 $M_\odot$	84.7 $M_\odot$	MS	3,700	0.03
5.0445	BH	35.1 $M_\odot$	82.2 $M_\odot$	CHeB	3,780	0.03
5.0445	BH	36.5 $M_\odot$	36.8 $M_\odot$	He star	43.8	0.00
5.3483	BH	36.5 $M_\odot$	34.2 $M_\odot$	He star	45.3	0.00
5.3483	BH	36.5 $M_\odot$	30.8 $M_\odot$	BH	47.8	0.05
10,294					0	0

- This left a binary black hole system with masses of $36.5\,M_\odot$ and $30.8\,M_\odot$, respectively, and orbital separation $a = 47.8\,R_\odot$.

- This system then spiraled together through gravitational wave emission for 10.3 billion years, merging about 1.1 billion years ago ($z \sim 0.09$) to produce GW150914.

The simulations described above paint a compelling picture but they entail large uncertainties because of

- *assumptions such as the metallicity*, and because
- accretion and (in particular) common envelope evolution are the *least-well understood aspects of binary evolution*.

Tests of these assumptions and increasingly strong constraints on models of massive binary star evolution may be expected as gravitational wave astronomy matures.

One crucial feature of the mechanism outlined above is *direct collapse of massive stars* in a binary to black holes,

- *without ejecting a supernova remnant* and
- without giving a strong *natal kick* to the black hole that is formed, so that it remains in the binary.

There is now some direct observational evidence that such *failed supernovae* may occur in nature, though we won't discuss it here.

22.5.2 Formation of Supermassive Black Holes

Is there a connection between the formation of stellar-size black holes and the formation of supermassive black holes found often in the centers of galaxies?

- Two pictures for the formation of supermassive black holes have been proposed.

  1. They may have formed by
     - successive merger of intermediate-mass black holes created by core collapse of massive first-generation stars, or
     - directly from the collapse of large clouds.
  2. The seeds for the growth of supermassive black holes may instead have been massive (say greater than $25\,M_\odot$) stellar black holes.

- In either case, it is possible that the evolution of massive stars leading to the creation of massive stellar black holes also has implications for the origin of supermassive black holes.

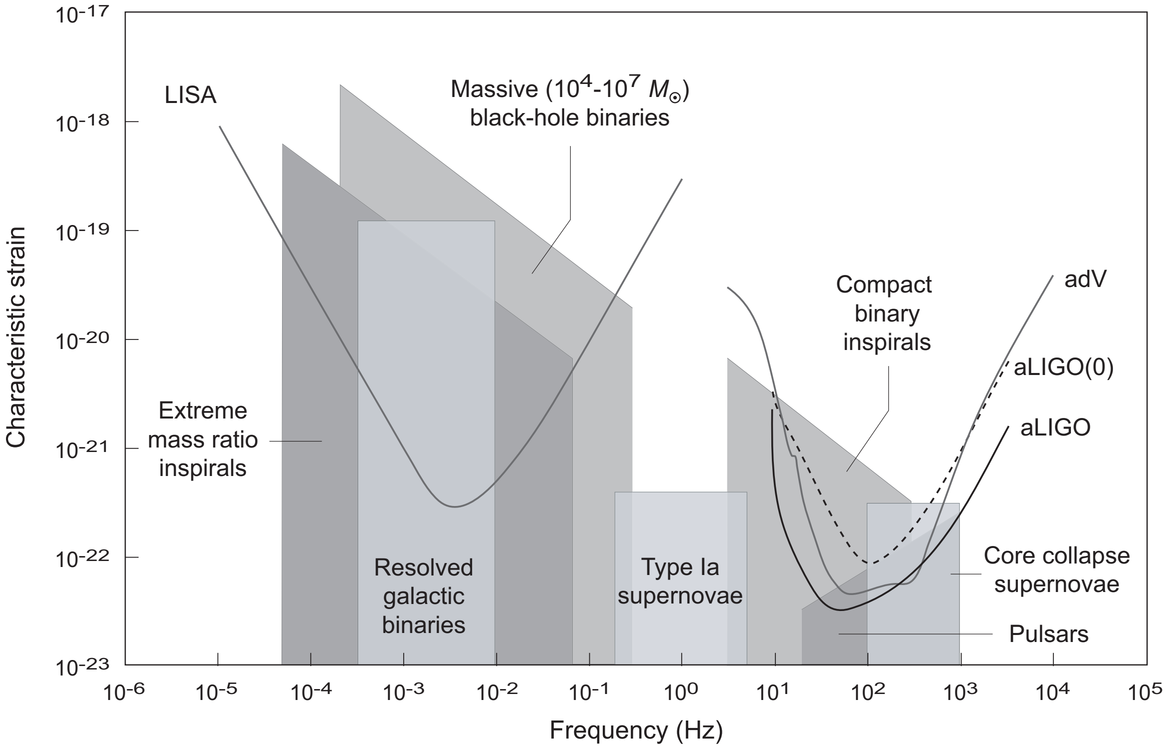


Figure 22.8: Strain amplitude and frequency ranges expected for gravitational waves from various astronomical sources. Minimum strain detection bounds for advanced LIGO (aLIGO) at full design capacity (∼2020), advanced Virgo (adV) at full design capability (∼ 2020), advanced LIGO in the first observing run after the upgrade [aLIGO(0), indicated by the dashed curve], during which the gravitational wave GW150914 was observed in 2015, and the proposed space-based array LISA are indicated.

Merger of supermassive black holes in galaxy collisions can't be studied with Earth-based observatories like LIGO because

- the gravitational wave frequency is too low, and
- the background noise level is too high,

But they could be studied in large space-based gravitational wave arrays; see Fig. 22.8.

22.6 Listening to Multiple Messengers

Prospects are good for the systematic accumulation of gravitational wave events from

- binary black hole mergers,
- binary neutron star mergers,
- mergers of neutron star–black hole binaries, and
- core collapse supernovae.

Even more interesting is the possibility of *multimessenger astronomy,* where, for example,

- a gamma-ray burst might be observed in coincidence with gravitational waves from a neutron star merger, or
- a neutrino burst might be observed in coincidence with gravitational waves from the accompanying supernova.

Since these events involve various aspects of late stellar evolution, multimessenger astronomy has the potential to revolutionize our understanding of how stars evolve.

In the following we shall summarize the *first gravitational-wave multimessenger event:* the coincidence of a gravitational wave with a gamma-ray burst and the subsequent electromagnetic transient.

22.7 Gravitational Waves from Neutron Star Mergers

On August 17, 2017, the LIGO–Virgo collaboration detected gravitational wave GW170817.

- This gravitational wave had a *very different signature* relative to previously-detected black hole merger events.
- The signal built slowly in amplitude and frequency with more than *3000 wave cycles recorded over almost 100 seconds* before peak.

This new kind of gravitational wave was quickly interpreted as originating in the *merger of two neutron stars*, but the show wasn't over yet!

- Approximately 1.7 seconds after the peak strain of the gravitational wave both the Fermi Gamma-ray Space Telescope (Fermi) and the International Gamma-Ray Astrophysics Laboratory (INTEGRAL) observed a *gamma-ray burst of two seconds duration* in the same part of the sky as the gravitational wave source.
- Within hours various observatories discovered a *new point source* in the irregular/elliptical galaxy NGC 4993 lying within the position error box for the gravitational wave.
- In the ensuing weeks a multitude of observatories studied the *transient afterglow in NGC 4993 (named officially AT 2017gfo)* intensively at various wavelengths.

Thus was the discipline of *multimessenger astronomy* born.

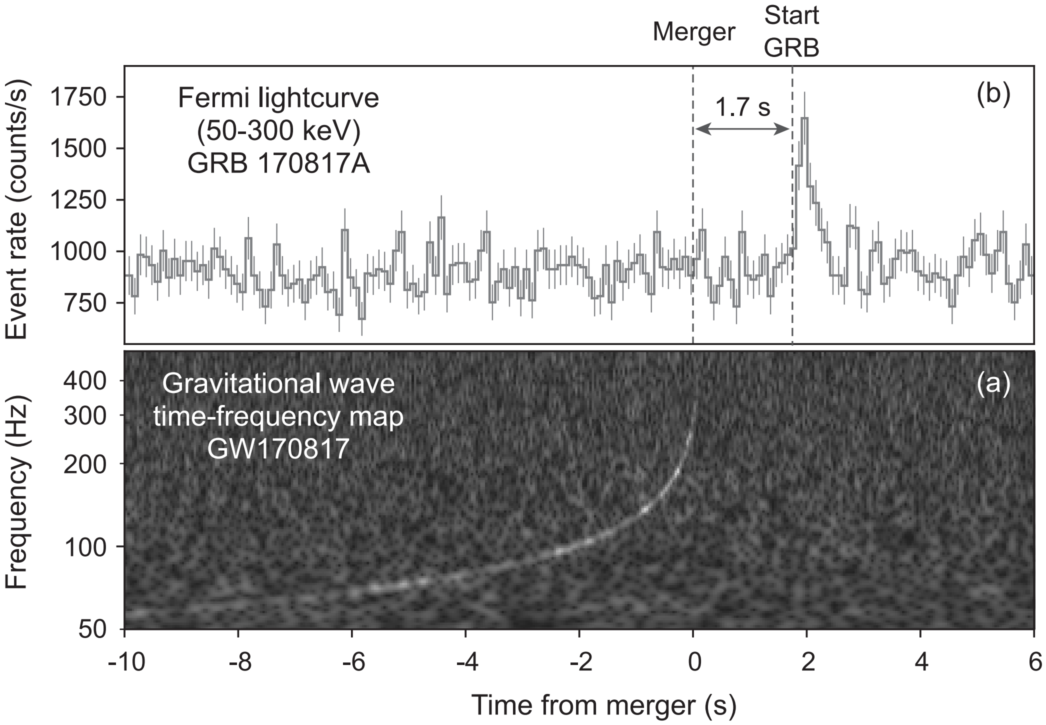


Figure 22.9: (a) Gravitational wave GW170817 (LIGO) and (b) gamma-ray burst GRB 170817A (Fermi satellite). The source was at a luminosity distance of 40 Mpc (130 Mly) and the gravitational wave and gamma-ray burst arrived at Earth separated by only 1.7 seconds.

The gravitational wave

- was identified by matched filtering against post-Newtonian waveform models and
- corresponded to the loudest gravitational wave signal observed to that date, with a *signal to noise ratio of 32.4*.

The coincidence of the gravitational wave and the gamma-ray burst is illustrated in Fig. 22.9.

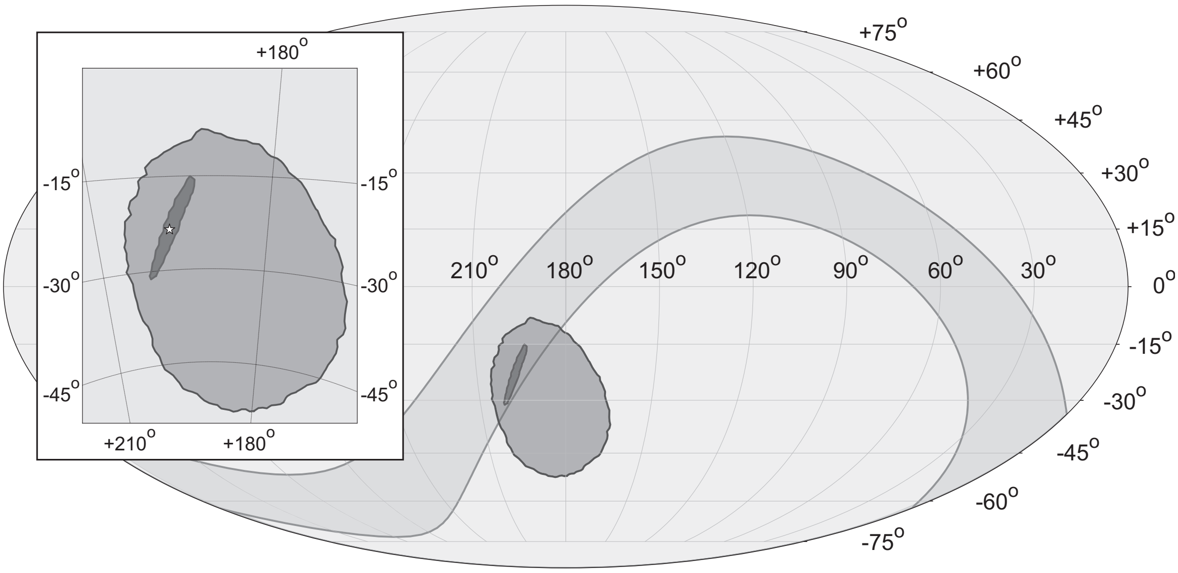
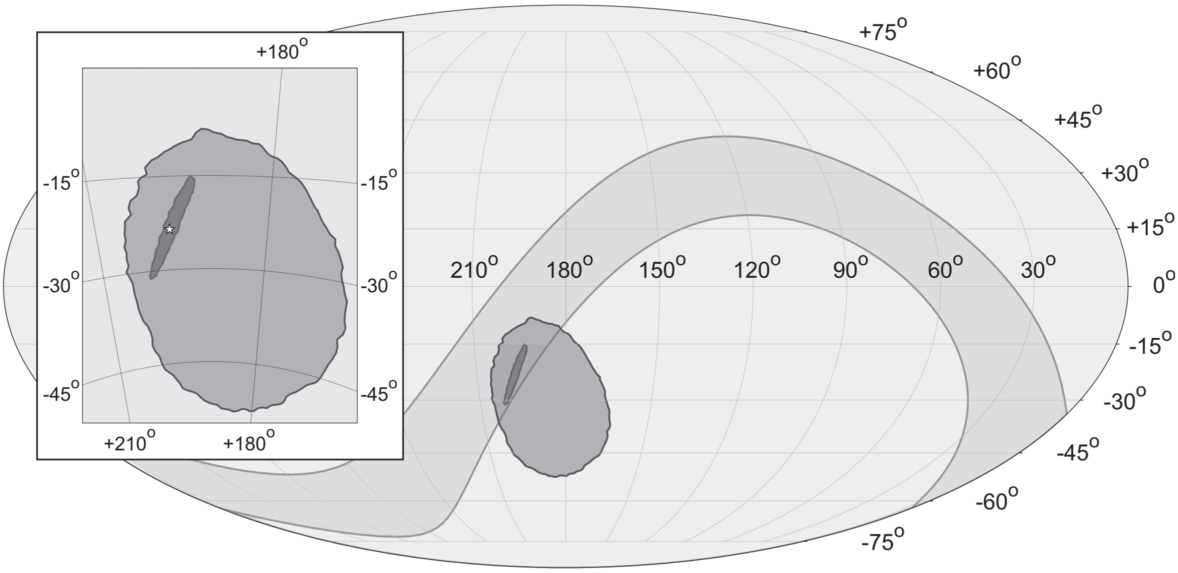


Figure 22.10: Localization of gravitational wave GW170817 and gamma-ray burst GRB 170817A. The 90% contour for LIGO–Virgo localization is shown in the darkest gray. The 90% localization for the gamma-ray burst is shown in intermediate gray. The 90% annulus from triangulation using the difference in GRB arrival time for Fermi and INTEGRAL is the lighter gray band. The zoomed inset shows the location of the transient AT 2017gfo (small white star) that was observed at various wavelengths. Axes correspond to right ascension and declination.

Sky localization of GW170817 is illustrated in Fig. 22.10.

- The final combined LIGO–Virgo sky position localization corresponded to an uncertainty area of 28 deg^2.
- The total mass determined for the binary was between $2.73\,M_\odot$ and $3.29\,M_\odot$, and the two individual masses were in the range $0.86\,M_\odot$ to $2.26\,M_\odot$.

These masses and the waveform indicate that the *two compact objects that merged were neutron stars*.



22.7.1 New Discoveries Associated with GW170817

The location of the afterglow is indicated by the small white star in the error box of the figure above.

- The luminosity distance was 40^{+8}_{-14} Mpc.
- Consistent with the known distance to the host galaxy.

The multimessenger nature of GW170817 proved to be a *treasure trove of discoveries* having fundamental importance in

- astrophysics,
- the physics of dense matter,
- gravitation, and
- cosmology.

Viability of Multimessenger Astronomy:

The event confirmed that the gravitational wave detectors could see and distinguish events that did not correspond to merger of two black holes.

- It also demonstrated for the first time that electromagnetic signals could be detected in coincidence with a confirmed gravitational wave event, and
- demonstrated sufficient source localization that the event could be observed at many different wavelengths.
- All told, more than 70 facilities observed the event at optical, radio, X-ray, gamma-ray, infrared, and ultraviolet wavelengths.

Mechanism for Short-Period GRBs:

- The interpretation of the event as the merger of binary neutron stars and
- the coincident (short-period) gamma-ray burst

provided the first conclusive evidence for the hypothesis that short-period gamma-ray bursts are produced in the merger of neutron stars.

- The gamma-ray burst was relatively weak, suggesting that the gamma-ray burst beam was not pointed directly at Earth.
- Confirmation of this hypothesis came two weeks after the initial event when radio waves and X-rays characteristic of a gamma-ray burst were detected.

This evidence taken together represents the first definite association of a gamma-ray burst with a progenitor.

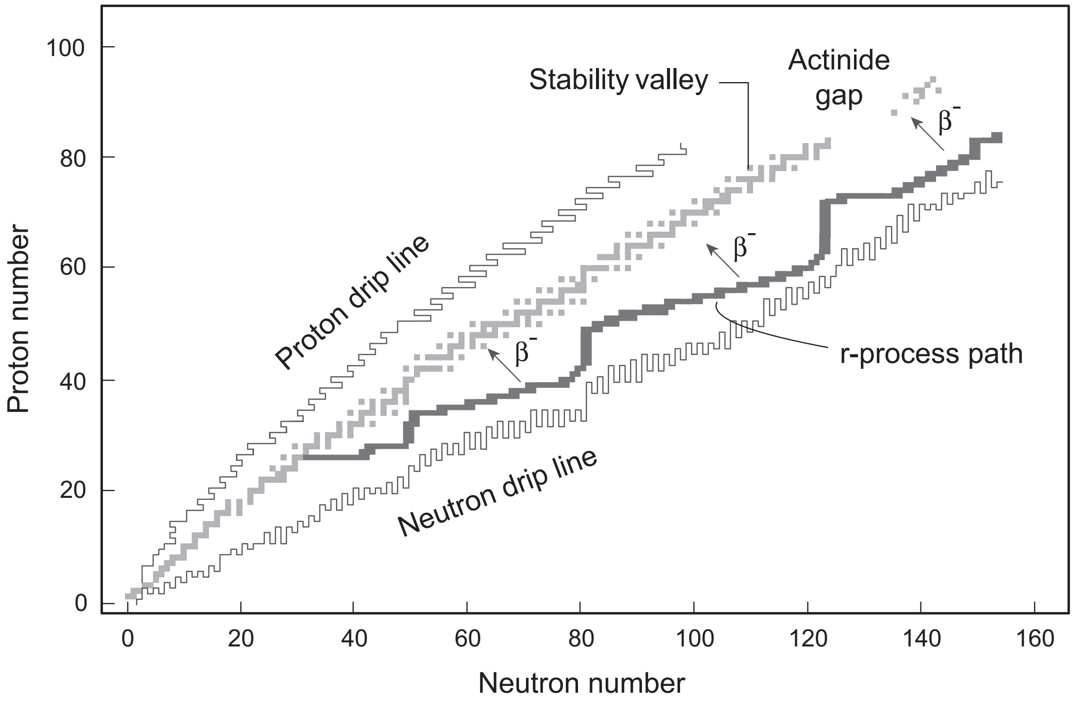
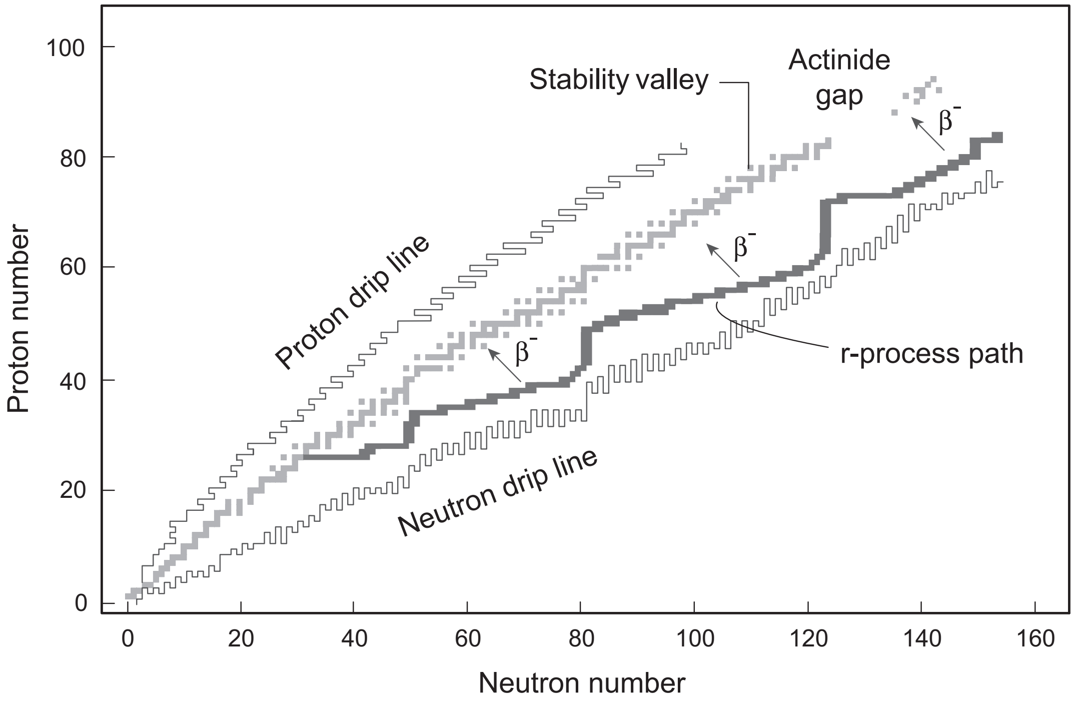


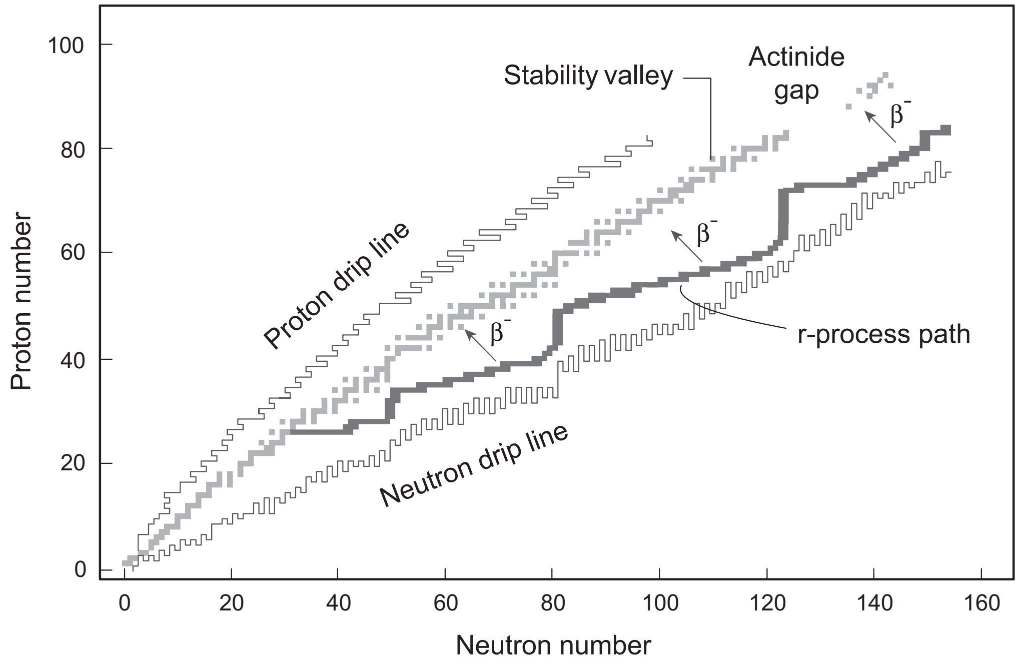
Figure 22.11: Theoretical path for the r-process. Nuclei produced along the r-process path will undergo rapid β^- decay back toward the stability valley, thus producing most of the neutron-rich and some of the β-stable isotopes, as well as all the actinide nuclei found in nature. (The β-stable isotopes beyond iron but below the actinide gap can be produced also in the slow neutron capture or s-process in red giant stars.) The two drip lines denote the boundaries beyond which a nucleus becomes unstable against spontaneous emission of neutrons or protons, respectively.

Site of the r-Process:

The signature of heavy-element production in the event demonstrated that neutron star mergers are one (perhaps the dominant) source of the rapid neutron capture or r-process thought to make many of the heavy elements (see Fig. 22.11).



- Now we have a quantitative way to investigate the relative importance of the two primary candidate sites for the r-process:
  - core collapse supernovae, and
  - neutron star mergers.
- Already it is clear that the dominant attitude of not very long ago that the r-process was associated mostly with core-collapse supernovae is probably not correct.



One common theme for understanding the origin of r-process nuclei is to ask whether they were produced

- *in a few rare events* (neutron star mergers occur maybe only once every million years in a large galaxy), or
- *in many much more common events* (core collapse supernovae occur about once every 50 years in a large galaxy).

Some evidence had been accumulating that at least some r-process nuclei were produced in rare events.

The neutron star merger leading to GW170817 gives direct evidence for *significant production of r-process nuclei in a single rare event*.

Observation of a Kilonova:

The expanding radioactive debris was observed at UV, optical, and IR wavelengths.

- This gave the first direct evidence for the *kilonova* (also termed a *macronova*) predicted to occur following such mergers as a result of radioactive heating by newly-synthesized r-process nuclei.
- The direct nucleosynthesis of r-process species likely ceases after a second or two, but most initially-synthesized isotopes would be highly radioactive and
- the cloud of debris can be kept warm ($10^3 - 10^4$ K) by radioactive decay for as long as weeks.

That the gamma-ray burst was emitted off-axis may have been essential in allowing the kilonova associated with the radioactivity of heavy elements produced in the merger to be observed.

As illustrated in Fig. 22.12 (following page), if the GRB is seen nearly on-axis, the GRB afterglow (“GRB transient”) masks the kilonova.

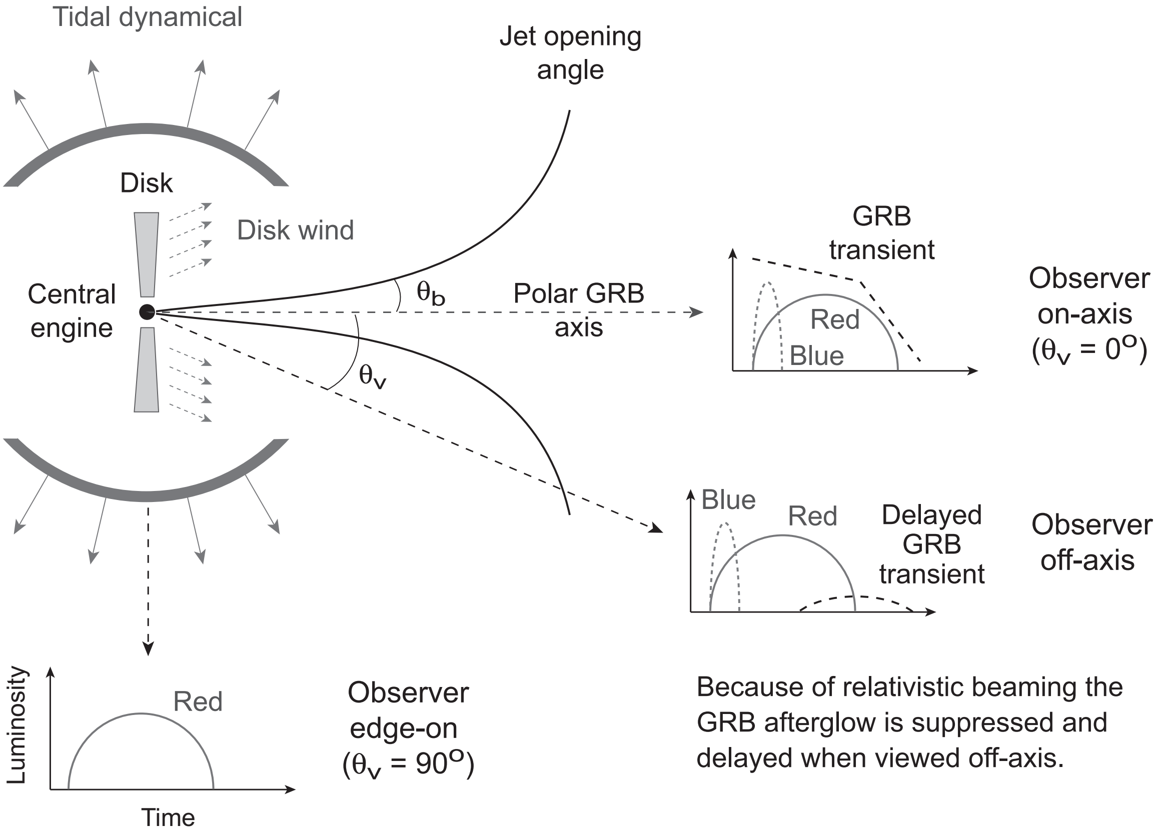
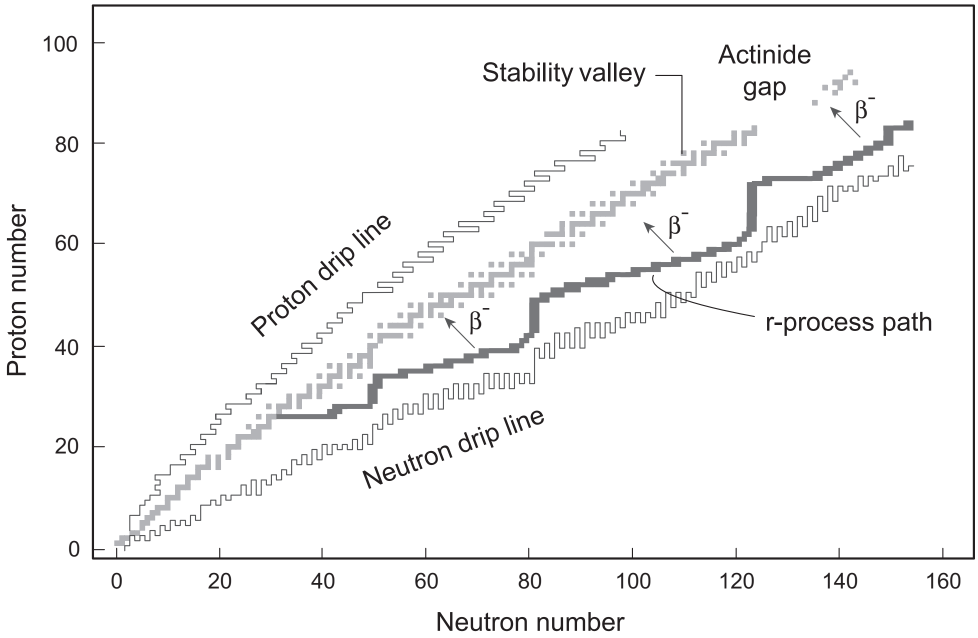


Figure 22.12: Geometry of GW170817 afterglows. Neutron-rich ejected matter labeled "Tidal dynamical" emits a kilonova peaking in the IR (solid arrows and solid curves labeled "Red" in the time–luminosity diagrams) associated with production of heavy r-process nuclei and high opacity (the *red kilonova*). Additional mass is emitted by winds along the polar axis (dotted arrows and dotted curves labeled "Blue") that is processed by neutrinos emitted from the hot central engine, giving matter less rich in neutrons and a kilonova peaking in the optical that is associated with production of light r-process nuclides and lower opacity (the *blue kilonova*). The usual GRB afterglow is indicated by dashed curves in the plots. It dominates all other emission when viewed on-axis but when viewed off-axis it appears as a low-luminosity component delayed by days or weeks (until $\theta_{\rm v} < \theta_{\rm b}$), which permits the kilonova to be seen.



Nuclei Far from Stability:

The r-process runs far to the right (neutron-rich) side of the β-stability valley in the chart of the isotopes shown above

- Little definitive information exists here because the isotopes cannot be made in traditional accelerators.
- Kilonova lightcurves are a statistical mix of contributions from many neutron-rich nuclei with no sharp lines because of the high velocities ($\sim 0.3c$) for the ejecta.
- However, they carry information about the average decay rates and other general properties of these largely unknown r-process nuclei.

This could provide future constraints on theories of nuclear structure far from β stability.

The Speed of Gravity:

The GRB arrived within 1.7 seconds of the gravitational wave from a distance of 40 Mpc.

- This established conclusively that the difference of the speed of gravity and the speed of light lies between -3×10^{-15} and $+7 \times 10^{-16}$ times c.
- (That is, no larger than 3 parts in 10^{15}).

Thus it took 1.7 seconds of observation to *eliminate from contention* theoretical alternatives to general relativity for which gravity does not propagate at c.

Neutron-Star Equation of State:

The multimessenger nature of the event indicates that neutron star mergers will provide an opportunity to make much more precise statements about the neutron-star equation of state.

- For example, the merger wave signature is *sensitive to the tidal deformability* of the neutron star matter near merger.
- This is of fundamental importance for our understanding of dense matter because prior observations have been unable to constrain candidate equations of state sufficiently to understand (for example)
  - the maximum mass of a neutron star and
  - the minimum mass of a black hole

  to better than an uncertainty of about a solar mass.

Demographics of Neutron-Star Binaries:

The observation of GW170817 provides quantitative information about the probability that neutron star binaries form in orbits that can lead to *merger in a Hubble time* .

- This probability has been rather uncertain to this point.
- The rate currently inferred corresponds to 0.8×10^{-5} mergers per year in a galaxy the size of the Milky way.

An accurate determination of the merger rate has implications for

- our *understanding of stellar evolution*,
- the *site of the r-process*, and
- the expected *rate of gravitational wave detection* from such events.

Determination of the Hubble Constant:

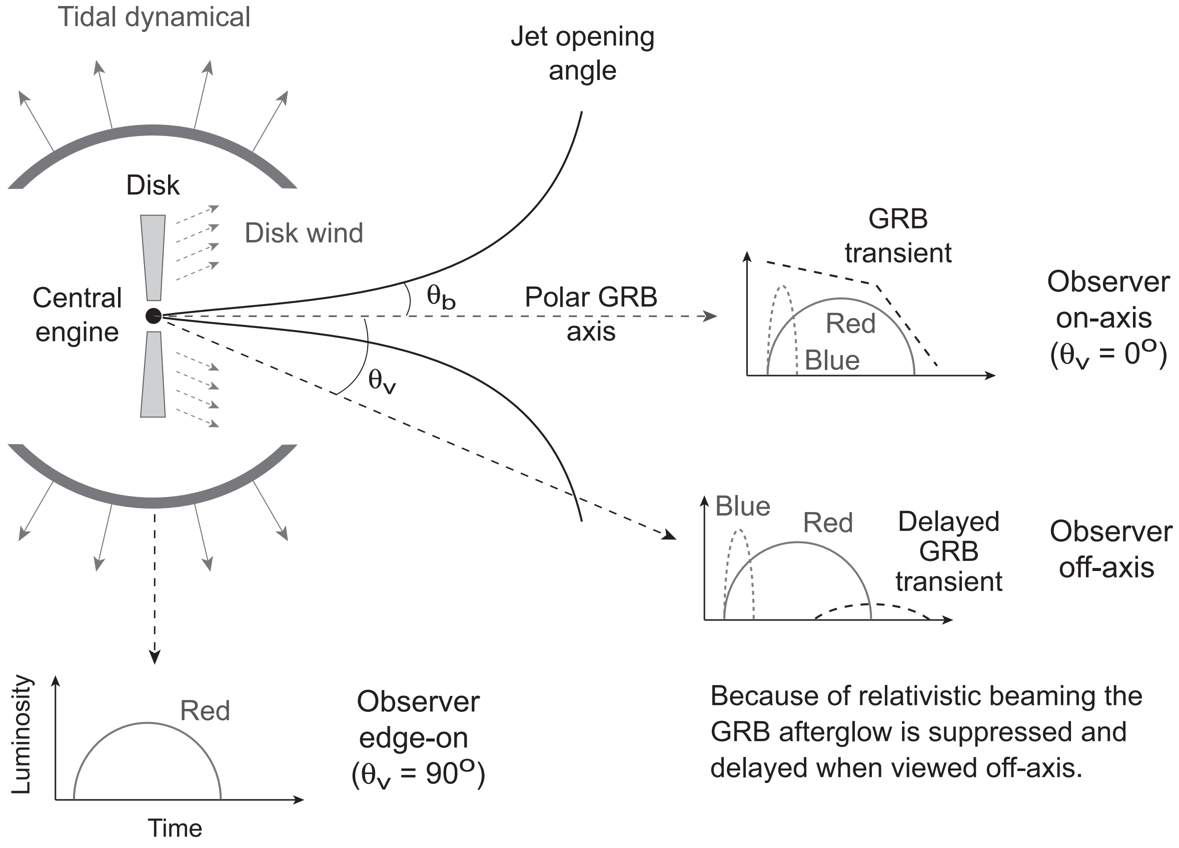
The multimessenger nature of the GW170817 event provides an independent way to determine the Hubble constant H_0.

- This can be accomplished by comparing the distance inferred from the gravitational wave signal with the redshift of the electromagnetic signal.
- Presently, different methods of determining H_0 yield a value in the range of about $67-73\,\mathrm{km\,s^{-1}Mpc}^{-1}$, with
  - analyses of the CMB tending to give values nearer the lower end and
  - traditional “distance-ladder” methods like Cepheid variables giving values nearer the higher end.
- Analysis of the GW170817 multimessenger event suggests a value in the middle,

$$H_0 \sim 70^{+12}_{-8}\,\mathrm{km\,s^{-1}Mpc}^{-1},$$

  but with relatively large uncertainties at this point.
- We may expect an accumulation of such multimessenger events to yield a *precise, independent determination of* H_0.

For example, 100 independent GW detections with host galaxy identified as in GW170817 could *determine* H_0 *with an uncertainty of* 5%.



Because of relativistic beaming the GRB afterglow is suppressed and delayed when viewed off-axis.

Off-Axis Gamma-Ray Bursts:

The initial observation of the kilonova followed days later by observation of X-ray and radio emission

- provides strong corroborating evidence for the *beamed nature of gamma-ray bursts* and
- represents the first clear detection of a *weak, off-axis GRB* and its slowing in the interstellar medium.

Systematic studies of such events should greatly enrich our understanding of gamma-ray bursts, which previously were understood in terms of bursts beamed more directly at us.

22.7.2 The Kilonova

Let's elaborate further on the *kilonova* powered by the *production of radioactive r-process nuclei*. GR simulations of neutron star mergers identify two mechanisms for mass ejection:

1. Matter may be
   - *expelled dynamically by tidal forces on millisecond timescales* during the merger itself, and
   - as surfaces come into contact *shock heating at the surfaces* may squeeze matter into the polar regions.
2. On a longer ($\sim 1\,\mathrm{s}$) timescale matter in an accretion disk around the merged objects can be *blown away by winds*.

As ejected matter decompresses, heavy elements are made.

- If the matter is highly neutron-rich, repeated neutron captures form the *heavy r-process nuclei* ($58 \leq Z \leq 90$);
- if the ejecta is less neutron-rich, *light r-process nuclei* ($28 \leq Z \leq 58$) are synthesized.
- Matter ejected in the tidal tails is cold and very neutron rich, and tends to form *heavy r-process nuclei*.
- The disk winds and ejecta squeezed into the polar regions
  - are *irradiated by neutrinos* from the central region, which *converts some neutrons to protons*.
  - This *favors the light r-process*.

The *photon opacity of the r-process ejecta* may play a central role in the observable characteristics of kilonova events.

- The photon opacity is generated largely by transitions between bound atomic states (*bound–bound transitions*).
- For light r-process nuclei the *valence electrons typically fill atomic d shells*.
- In contrast a substantial fraction of heavy r-process species produced by simulations (often 1–10% by mass) are *lanthanides* ($58 \le Z \le 71$).
- For lanthanides the *valence electrons fill the f shells*.
- These have densely-spaced energy levels and an *order of magnitude more line transitions* than for the d shells in light r-process species.
- As a consequence,

  > *The opacity of heavy r-process nuclei is* roughly a factor of 10 larger *than the opacities for light r-process species*,

  and they have correspondingly long photon diffusion times.

Hence the cloud of *light r-process species*

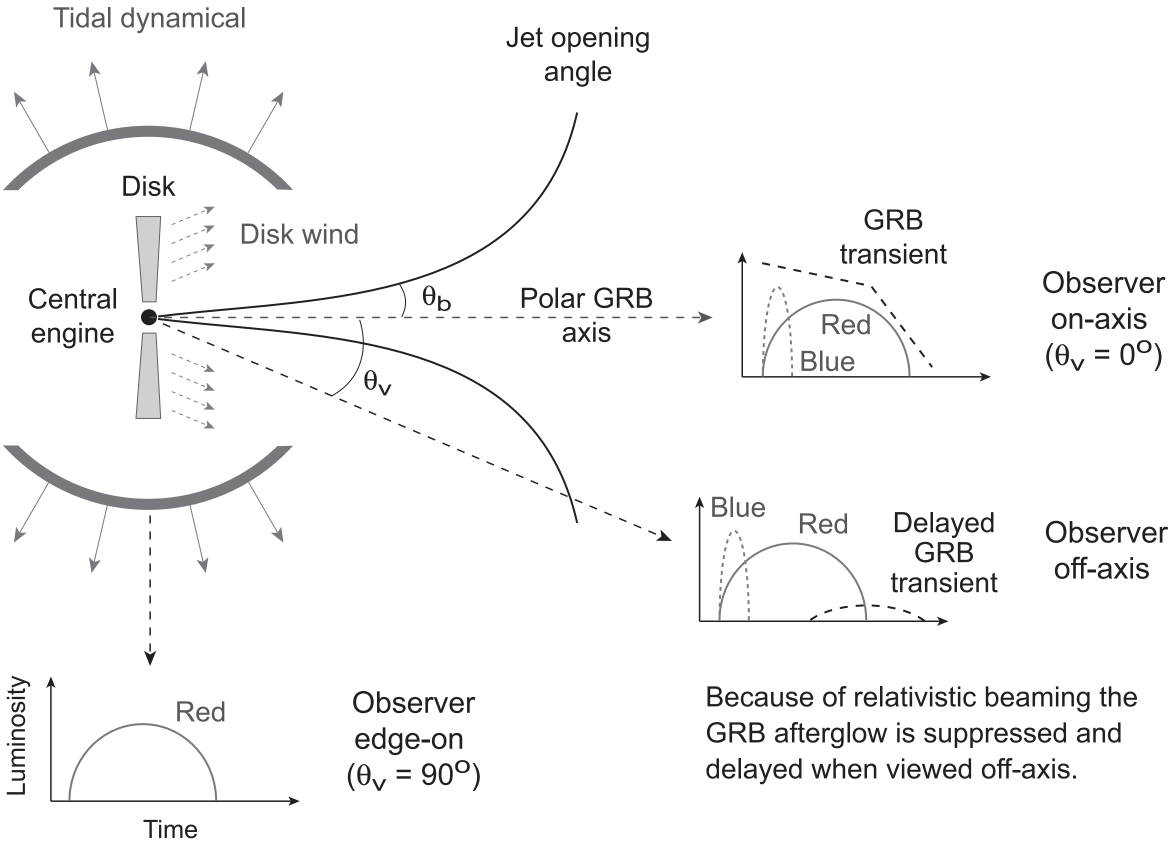
- is considerably *less opaque*
- has *shorter diffusion times*, and
- *tends to radiate in the optical* and fade over a matter of days.

In contrast, the cloud of *heavy r-process species*

- radiates *in the IR*
- *for as long as weeks* because of the
  - *high opacity* and
  - long *diffusion times*.

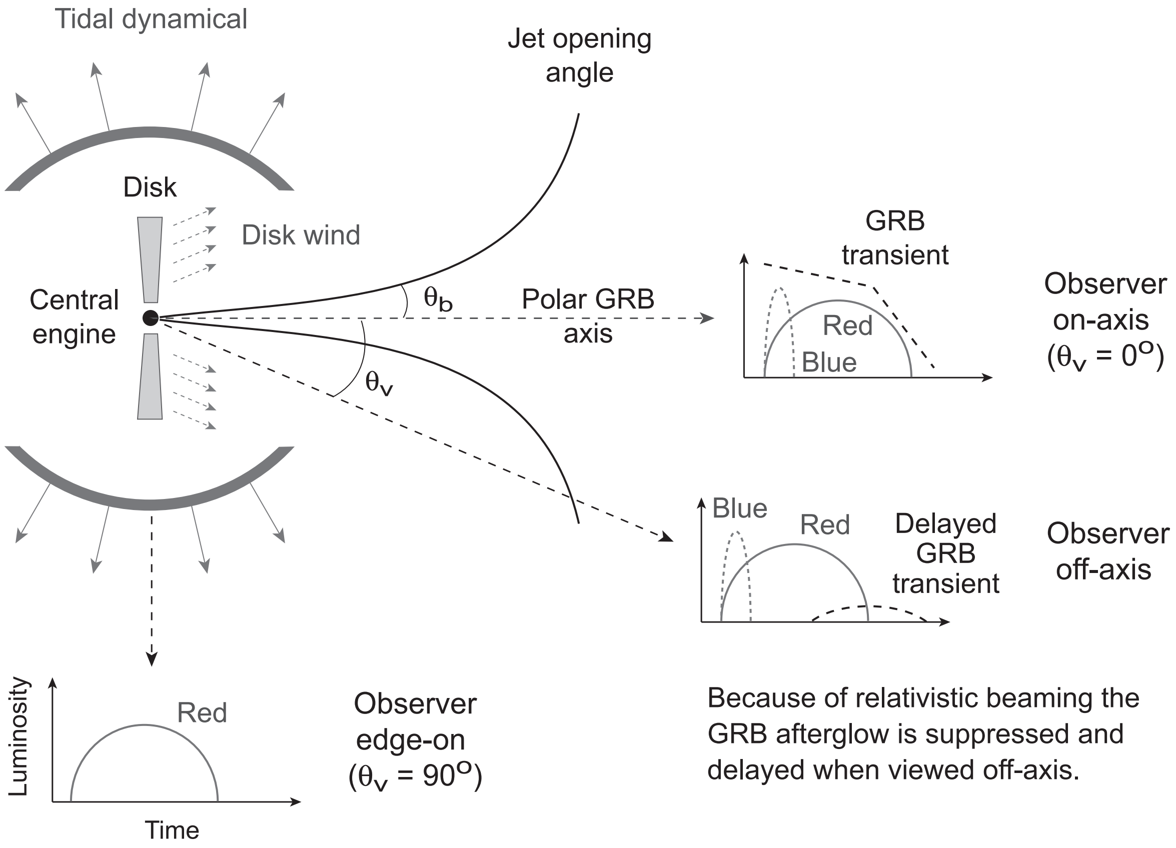
This accounts for the observed characteristics of the transient AT 2017gfo, which *differed essentially* from all other previous astrophysical transients:

- It *brightened quickly in the optical* and then faded but
- a quickly-growing *IR emission* remained strong for many days, and
- only after a period of weeks did *X-ray and RF signals* begin to emerge.



The preceding considerations suggest a general picture of the geometry of GW170817 that is illustrated in the figure above.

- The kilonova transient AT 2017gfo that followed the gravitational wave GW170817 and associated gamma-ray burst GRB 170817A had two distinct components.
- First, *tidal dynamical ejection* flung out on ms timescales very neutron-rich matter at high velocities $v \sim 0.3c$.
  - This matter underwent extensive neutron capture to produce *heavy r-process species*.
  - It had *high opacity* because of the lanthanide content.



Because of relativistic beaming the GRB afterglow is suppressed and delayed when viewed off-axis.

- Secondly, *winds ejected matter from the disk region* on a timescale of seconds.
  - This matter was subject to *shock heating and to irradiation by neutrinos* from the hot center.
  - Both tended to *decrease the neutron to proton ratio*.

Nucleosynthesis in this less neutron-rich matter was likely to produce *light r-process matter of lower opacity*, since there weren't enough neutrons to produce lanthanides and other heavy r-process nuclei.

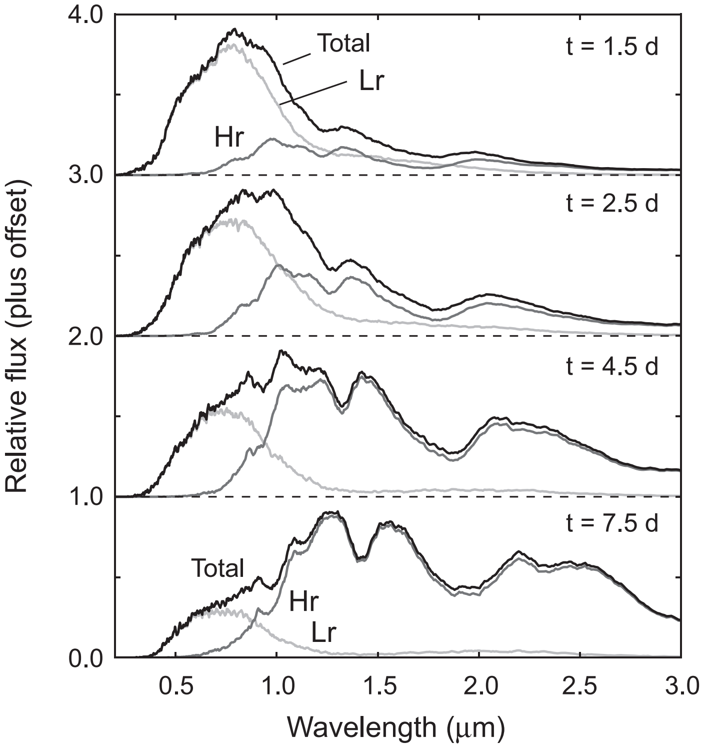
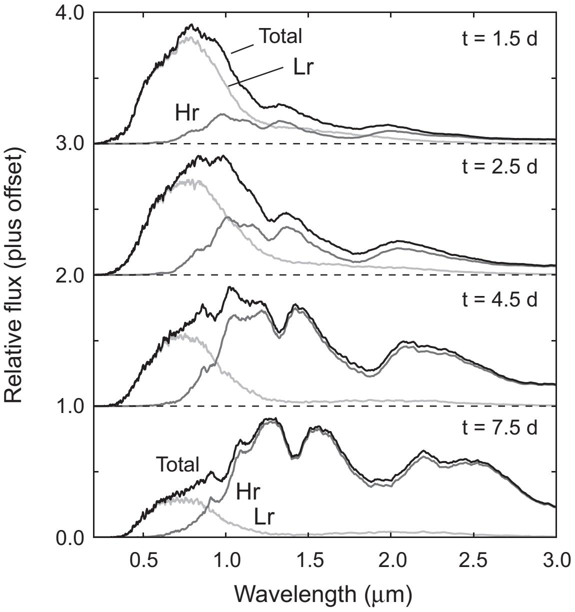


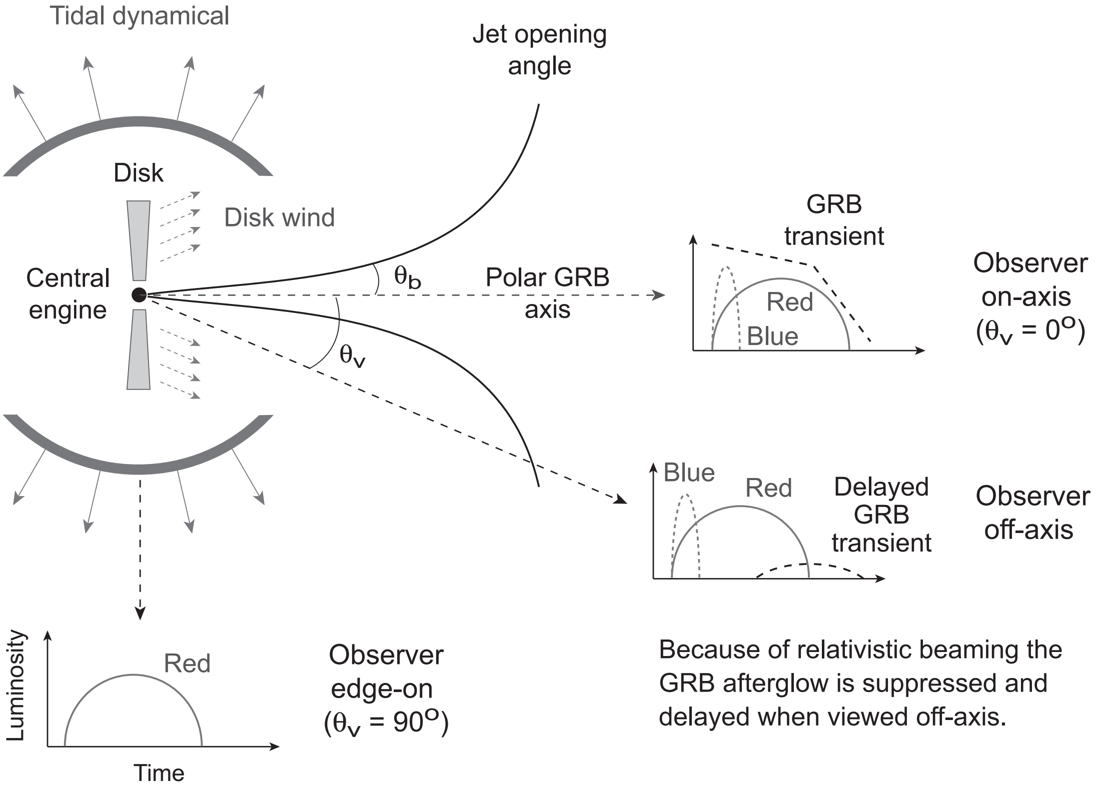
Figure 22.13: Evolution of GW170817 kilonova components. Total flux is a sum of two spatially-separated components: dominantly-optical emission from light r-process isotopes ("blue kilonova", labeled Lr) and dominantly-IR emission from heavy r-process isotopes ("red kilonova", labeled Hr).

This picture is supported by simulations in Fig. 22.13.

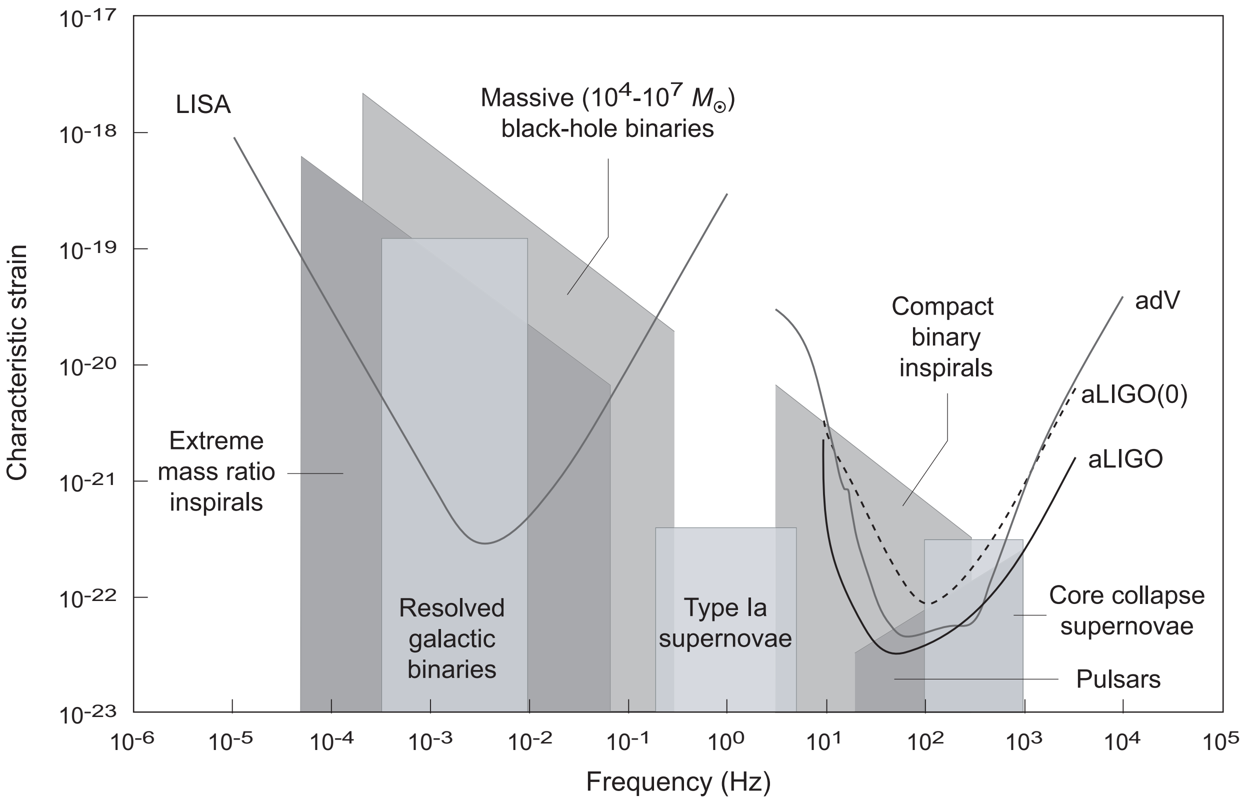
- These simulations exhibit clearly the early emergence and rapid decay of the optical component associated with the light r-process (the *blue kilonova*).
- This is followed by the longer-lived IR component associated with the heavy-r process (the *red kilonova*), which grows within several days to dominate the lightcurve.



- Red and blue components are visible only because the *GRB afterglow was suppressed by relativistic beaming*.



Because of relativistic beaming the GRB afterglow is suppressed and delayed when viewed off-axis.

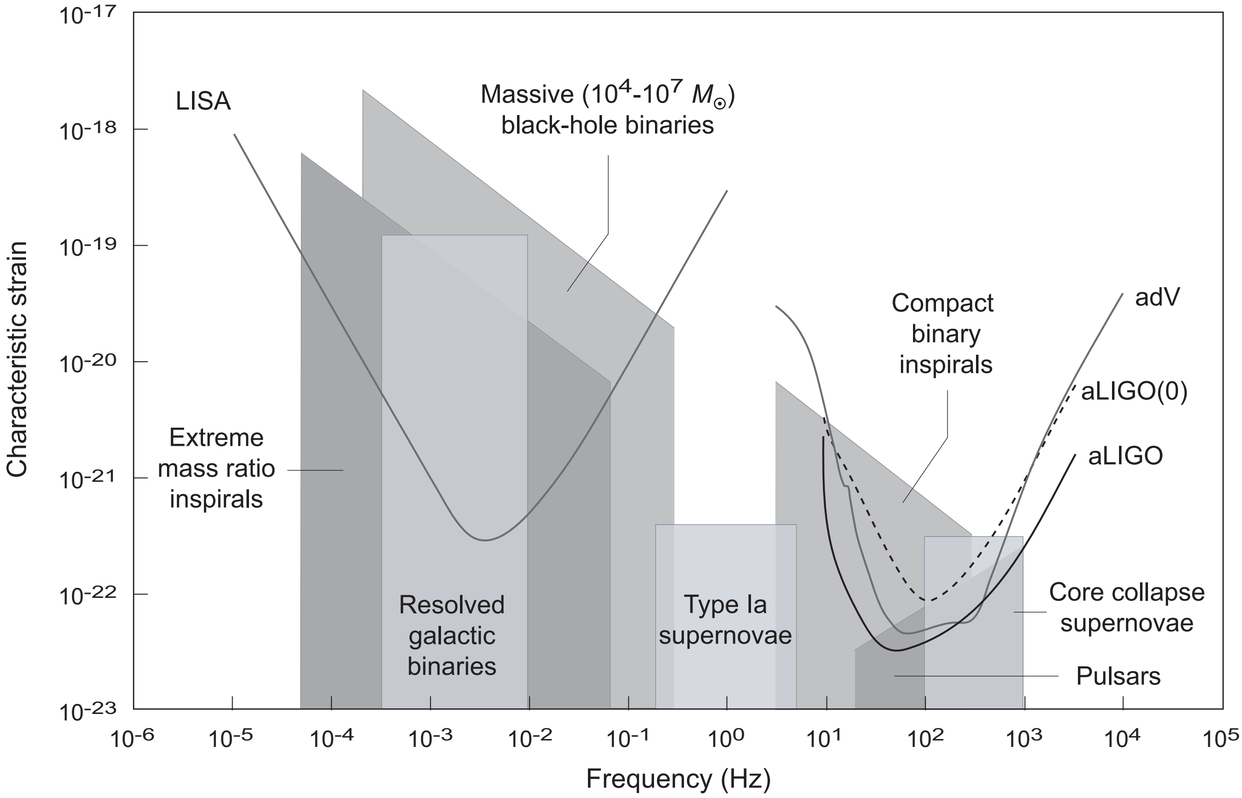


22.8 Gravitational Wave Sources and Detectors

Let’s conclude with an overview of the prospects for detecting gravitational waves from various astrophysical sources.

- Amplitude and frequency ranges for operating and proposed gravitational wave observatories, along with
- corresponding ranges expected for some important astrophysical sources of gravitational waves,

are reproduced above.



- Earth-based detectors like LIGO and Virgo are prime instruments for elucidating the physics of
  - neutron stars,
  - black holes, and
  - core collapse supernovae.
- Space-based arrays could probe gravitational waves from
  - merger of supermassive black holes in galaxy collisions, and
  - ordinary binary stars in the galaxy.